UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Trabajo Fin de Grado

# Detección de la Saliencia Auditiva en Registros de Audio

Autora: Miriam Serrano Salinas

Tutora: Carmen Peláez Moreno

Marzo 2017

## Resumen

La percepción humana es un proceso por el cual nuestro cerebro recibe información a través de los sentidos del mundo que nos rodea. Sin embargo, durante este proceso, algunos estímulos son considerados más importantes que otros, es decir, se priorizan.

La saliencia auditiva define, por tanto, el mecanismo que utiliza nuestro cerebro para priorizar ciertos estímulos, en este caso de tipo sonoro.

Durante los últimos años, los avances tecnológicos y la adaptación de modelos para saliencia visual, han supuesto el comienzo definitivo de la investigación en el campo de la detección de eventos auditivos salientes.

Además, el entrenamiento de redes neuronales para su aplicación en estos modelos permite obtener una aproximación más cercana a la estructura biológica real que genera el proceso de priorización.

Diversos tipos de redes neuronales son implementados en función del objetivo del modelo desarrollado. En algunos casos, la finalidad será clasificar eventos, en otros la detección. Para el caso de este proyecto, se utiliza la regresión como modelo para obtener valores numéricos que permitan ajustar los pesos de la red neuronal en función de los valores objetivo, obtenidos mediante mediciones fisiológicas para formar un *ground truth*, es decir, un valor fiable de referencia.

En los últimos años, ya están surgiendo modelos más complejos que comprenden la detección de saliencia auditiva y visual conjuntamente, ya que en ámbitos como el cinematográfico o incluso en nuestra vida diaria es más natural utilizar ambos sentidos, el de la vista y el del oído, de manera combinada.



**Palabras clave: Saliencia auditiva, redes neuronales, regresión, ground truth.**

## Abstract

Human perception is a process that our brain receives information through the senses from the world around us. However, during this process, some stimuli are considered more important than the others, i.e, they are prioritized.

Aural saliency defines the mechanism that our brain use to prioritize certain stimuli, in this case sounds.

During the latest years, the technology advances and the adaptation of models for visual saliency, have been the beginning of the aural salience event detection research.

Furthermore, the neural network training for the application in these models let us to obtain an approach to the biological structure that generates the priority process.

Several neural networks types are implemented depending on the objective of the model developed. In some cases, the finality will be the event classification, other times the detection. In this project, we use the regression model to obtain number values that allow adjust the weights of the neural network in accordance with the objective values, which are obtain through physiological measurements to form the ground truth, i.e., the reference.

In this years, more complex models are emerging. This models include de aural and visual saliency because some contexts as the cinema or even the daily life is more natural to use both senses, the sense of sight and hearing combined.



**Key words: aural saliency, neural network, regression, ground truth.**

# Índice General

## Índice de figuras

# Índice de Tablas

# Chapter 1

# Introduction

## *1.1. Context*

The human brain receives an anormous of information that is filtered through mechanisms that prioritizes some stimuli over others, this is called saliency.

Since these stimuli are perceived by different senses, we can distinguish between different modalities. In particular, in this Project we are interested in the auditory modality and we will focus on the mechanisms that makes us focus our attention on certain auditory stimuli over others of its class.

In this field, there are very few studies, very recent and an increasing interest to the scientific community and the industry is being observed. This foresees the emergence of new tools and research related to it in the coming years.

On the other hand, to measure physiological response, new devices, cheaper and easier to use, have appeared providing new studies in this field with a more reliable ground truth.

This Project proposes an auditory salience detection model based on the use of a number of related descriptors, using audio files from films under *Creative Commons* database LIRIS-ACCEDE and *Galvanic Skin Response* (GSR) annotations.

## *1.2. Objectives*

Saliency detection is attracting increasing interest from the scientific community. Within this context, the aural saliency is not very developed yet. Instead, some studies have adapted the visual model to auditory characteristics, but this is not very satisfactory yet.

During the last years, some researchers have published new specific databases for particular cases but it is not easy to find general purpose results.

To develop this project, we will use an experiment based on physiological measurements, musical descriptors and several videos that include audio to approximate a general model to predict the saliency for every audio input, able to provide an appropriate response by training an artificial neural network, using LIRIS-ACCEDE database that has the advantage of providing open data to avoid copyright problems.

On the other hand, the integration of this model into a multimodal model, i.e., visual and aural models together, is proposed as a future work due to the relationship between aural and visual stimuli in many cases.

## *1.3. Stages of development*

This Project is organized into the following stages:

1. Research and documentation:
    a. Research process on existing projects in this area.
    b. Searching descriptors to detect auditory saliency.
2. Software development:
    a. Choosing appropriate libraries.
    b. Creating the code in MATLAB.
        i. MATLAB code development.
        ii. Training the Neural Network.
3. Running the software:
    a. Setting parameters for optimum results.
4. Result analysis and writing the documentation:
    a. Analysis of results compared with baseline.
    b. Writing the paper based on the results.

For further information, some details are given below:

During the process of finding work about the audio salience detection, the way the brain decodes and processes information registered through the sense of hearing has been investigated.

Later, research projects related to the detection of human emotions using algorithms based on machine learning were compiled. In this case, the dataset used comes from LIRIS-ACCEDE [1].

## *1.4. Document structure*

This paper is divided into 6 chapters in which the first is the introduction with paragraphs for the context of the topic, the goal of the research, phases of this project, the structure of this document, socio-economic environment and regulatory framework.

In the second chapter, the State of the Art is reviewed, emphasizing some research on auditory salience and explaining differences between bottom-up saliency and top-down detection.

In chapter three, the developed system to obtain descriptors that let us compute audio saliency with MATLAB software is presented.

In chapter four, the experiments conducted and results obtained for each file in the database used are presented.

In chapter five, we summarize the conclusions reached with these results and the future work is proposed.

In chapter six, a budget estimated by calculating the time and money invested in this Project is shown in detailed tables.

An appendix is added with a summary of the project.

The last part of this document is another appendix to compile the graphical representations of some relevant parameters in the performance of the algorithm.

## *1.5. Socio-economic environment*

The aural saliency detection is very useful in some tasks like audio and acoustic event classification, detection of non-vocalization sounds in meetings or automatic speech recognition (ASR) [34], new tools that provide more facilities in human-machine interaction, for example.

Furthermore, recent studies in salient stimuli detection and event classification [35] often integrate auditory and visual saliency since humans tend to prioritize stimuli based on various models simultaneously, so that influences other applications like film industry, marketing, etc.

## *1.6. Regulatory framework*

In this project, two things are identified that can subject to regulation: the aural saliency model and physiological response measurements.

Neither of them have been legislated: in the firs case, aural saliency detection, is still a matter under study. In the second case, thought is has not been specifically regulated, physiological response is a private data and therefore it is included in the *data protection law* [36]. Thus, participants in the annotation of GSR measures had to sign an informed consent form.

# Capítulo 2

## *2.1 Medida de la Saliencia auditiva*

En el ámbito de la investigación, se diferencian dos modelos a la hora de referirse a la detección de la atención humana:

- Atención *bottom-up*: Se activa mediante un estímulo externo, de forma rápida e involuntaria.
- Atención *top-down*: Se desencadena cuando la persona detecta un acontecimiento que tiene interés para ella, por lo que se podría decir que tiene una componente más subjetiva. Ocurre de manera más lenta y premeditada.

Las primeras investigaciones sobre saliencia auditiva se basaban en el modelo de atención *bottom-up* y eran, básicamente, una adaptación del trabajo para la atención visual de Itti et al. [3]. En estos trabajos se calculan espectrogramas de audio pero se tratan como imágenes [4, 5].

Estas adaptaciones presentan ciertos inconvenientes:

- No se trata de la manera correcta la dimensión temporal del sonido.
- En muchos casos, las características de bajo nivel elegidas (frecuencia, contraste temporal, etc.) no son suficientemente representativas de la percepción auditiva humana.

Para evitar los problemas mencionados, se han propuesto otros modelos basados en estadísticos como el de T. Tsuchida y G. W. Cottrell [6] en 2012, o un año más tarde el de B. Schauerte y R. Stiefelhagen [7], que utilizan características de bajo nivel como: la envolvente de la forma de onda, el ancho de banda, tono, etc. apoyándose en el análisis basado en el banco de filtros Gammatone [6], más adecuadas que las de los primeros modelos.

Los avances en la detección de la atención *bottom-up* implican la incorporación de características de bajo nivel bio-inspiradas pretendiendo así imitar el sistema auditivo humano [8, 9] o la creación de nuevas parametrizaciones enfocadas a la clasificación de eventos acústicos [10, 11, 12, 13, 14, 15].

Para escenas acústicas complejas, la tendencia es la creación de modelos de saliencia auditiva *top-down* y *bottom-up* combinados, éste es el caso de las publicaciones de O. Kalinli y S. S. Norayanan [5] y B. D. Coensel y D Botteldooren [16].

En este proyecto, el modelo desarrollado se basa en atención *bottom-up* ya que la detección se realiza partiendo de la premisa de que la atención se activará a través de un estímulo externo que, en este caso, será de tipo auditivo sin intervención de motivaciones de nivel superior y no orientado a ninguna tarea en particular.

### 2.1.1 Medida del *ground truth*

Se apuntaba en la introducción que una de las dificultades de desarrollar un modelo para la detección de la saliencia auditiva era la obtención de un *ground truth* fiable dada la dificultad de medir, de manera objetiva, datos en el sujeto sobre estímulos sonoros.

En este sentido, han ido surgiendo varias alternativas:

- Pruebas donde el participante decide, de manera subjetiva, cuál ha sido el estímulo sobresaliente, ejemplo de este procedimiento sería el modelo de C. Kayser et al. [4] o el de T. Tsuchida y G.W. Cottrell [6].
- Igual que el procedimiento anterior, pero añadiendo a las anotaciones subjetivas del sujeto medidas objetivas de ciertos parámetros como el tiempo de respuesta, utilizado en el trabajo realizado por F. Tordini et al. [17].
- Etiquetado manual del audio donde el sujeto escucha grabaciones de escenarios reales y realiza las anotaciones a través de una interfaz.

Se han publicado durante estos años algunas bases de datos específicas siguiendo estas alternativas con el objetivo de obtener medidas fiables para generar un *ground truth* aplicable a estudios futuros. Sin embargo, la mayoría de las bases de datos públicas creadas para su uso en el aprendizaje máquina aún son, en su mayoría, poco realistas, de tamaño reducido y suelen tener problemas de copyright como apunta la publicación de Yoann Baveye et al. [19].

A continuación se muestran en una tabla las más relevantes surgidas en los últimos años:

| ***Nombre*** | ***Tamaño*** | ***Etiquetas*** |
|---|---|---|
| HUMAINE | *50 clips de entre 5 segundos y 3 minutos de duración.* | Etiquetas globales: estados emocionales, etiquetas de contexto, eventos clave, palabras relacionadas con emociones, etc.<br><br>Etiquetas a nivel de frame: intensidad, excitación, valencia, dominancia, previsibilidad, etc. |
| FilmStim | 70 extractos de películas de entre 1 minuto y 7 minutos de duración. | 24 criterios de clasificación: excitación subjetiva, afecto positivo y negativo, puntuaciones afectivas positivas y negativas derivadas de la Escala Diferencial de Emociones, seis puntuaciones afectivas discretas y 15 puntuaciones relativas a emociones mixtas. |
| DEAP | 120 vídeos musicales de 1 minuto de duración. | Clasificaciones mediante autoevaluaciones online de excitación, valencia y dominancia y grabaciones en vídeo de la expresión de la cara para un conjunto de 40 vídeos musicales. |
| MAHNOB-HCI | 20 extractos de películas de entre 35 y 117 segundos de duración. | Palabras clave relacionadas con emociones, excitación, valencia, dominancia y previsibilidad combinadas con vídeos de la expresión de la cara, |

| | | |
|---|---|---|
| | | EEG, audio, mirada y grabaciones fisiológicas periféricas. |
| EMDB | 52 clips de vídeo de 40 segundos de duración y sin audio. | Clasificaciones globales para la excitación, valencia y dominancia inducidas. |
| VIOLENT SCENES DATASET | 25 películas completas. | Incluye la lista de segmentos de película que contienen violencia física según dos definiciones diferentes y, también, 10 términos de alto nivel para las modalidades de audio y vídeo: presencia de sangre, peleas, balazos, gritos, etc. |
| LIRIS-ACCEDE | 9800 extractos, a partir de 160 películas, de entre 8 y 12 segundos. | Clasificaciones para las dimensiones de excitación y valencia. |

***Tabla 1.*** *Tabla con las diferentes bases de datos publicadas para detección de emociones [1].*

Como se puede ver en la tabla anterior, existen varias bases de datos disponibles para su utilización en modelos de detección de saliencia de tipo audiovisual. A continuación, se exponen las características más relevantes de cada una de ellas:

| ***Nombre*** | ***Características más relevantes*** |
|---|---|
| HUMAINE | Creada por Douglas-Cowie et al. [26].<br><br>Más de 50 clips de entre 5 segundos y 3 minutos de duración. |
| FilmStim | Creada por Shaefer et al. [27].<br><br>70 extractos de entre 1 y 7 minutos a partir de 10 películas que incluyen todas las categorías emocionales: enfado, tristeza, miedo, disgusto, aburrimiento, ternura y estado neutro.<br><br>364 participantes utilizando 24 criterios de clasificación. |
| DEAP | Creada por Koelstra et al. [28].<br><br>120 extractos de vídeos musicales de 1 minuto de duración.<br><br>Calificaciones de los extractos online mediante una autoevaluación de 14 voluntarios como mínimo para cada extracto.<br><br>El acceso a los vídeos se da mediante un enlace de YouTube. |
| MAHNOB-HCI | Creada por Soleymani et al. [29].<br><br>Base de datos multimodal con 20 extractos a partir de películas comerciales.<br><br>30 participantes. |
| EMDB | Creada por Carvalho et al. [30]. |

| | |
|---|---|
| | 52 clips de vídeo sin sonido extraídos de películas comerciales de 40 segundos de duración.<br><br>113 participantes para calificar valencia, excitación y dominancia inducidas en una escala de 9 puntos. |
| VIOLENT SCENES DATASET | Creada por Demarty et al. [31].<br><br>Se basa en la extracción de eventos violentos en películas.<br><br>Se proporcionan los enlaces a los DVDs usados para el proceso de anotación en el sitio web de Amazon. |
| LIRIS-ACCEDE | Utiliza el espacio valencia – excitación 2D. [1]<br><br>Está abierta a aportaciones que ayuden a ampliarla y diversificarla.<br><br>Comprende mediciones fisiológicas de 30 de las 160 películas que la forman para utilizarse como *ground truth.* |

***Tabla 2.*** *Tabla con las características de cada base de datos considerada en este apartado [1].*

La dificultad estriba en que la mayoría de estas bases de datos están dirigidas a un fin muy concreto para el que han sido creadas. Algunos estudios, incluso, reclaman la creación de una base de datos estándar para su aplicación en el ámbito de la computación afectiva [32, 33].

## *2.2 Descriptores para la extracción de características*

Una parte fundamental para la obtención de un modelo fiable que permita detectar estímulos salientes es elegir buenos descriptores que se ajusten correctamente a las características del evento. En este caso, como se ha mencionado anteriormente en este documento, no hay demasiados trabajos realizados en concreto para la saliencia de tipo auditivo pero, en general, los que se han publicado más recientemente, se basan en las siguientes características:

- La envolvente de la forma de onda.
- Tono.
- Ancho de banda.
- Parámetros relacionados con el análisis basado en banco de filtros Gammatone [6].
- Parámetros basados en centroides temporales y espectrales.

Existen, además, librerías diseñadas específicamente para el análisis de archivos musicales, como la *Toolbox* de MATLAB denominada *MIRToolbox* [20] que facilitan este tipo de análisis ya que comprenden multitud de funciones que calculan todos los parámetros que puedan resultar de interés para la detección de saliencia auditiva como pueden ser: *mirfilterbank* para la descomposición de la señal en frecuencia, *mirenvelope* para la envolvente de onda, *mirspectrum* para la descomposición de la energía de la señal, etc. En la siguiente figura se muestran las funciones que forman esta librería y sus interconexiones:

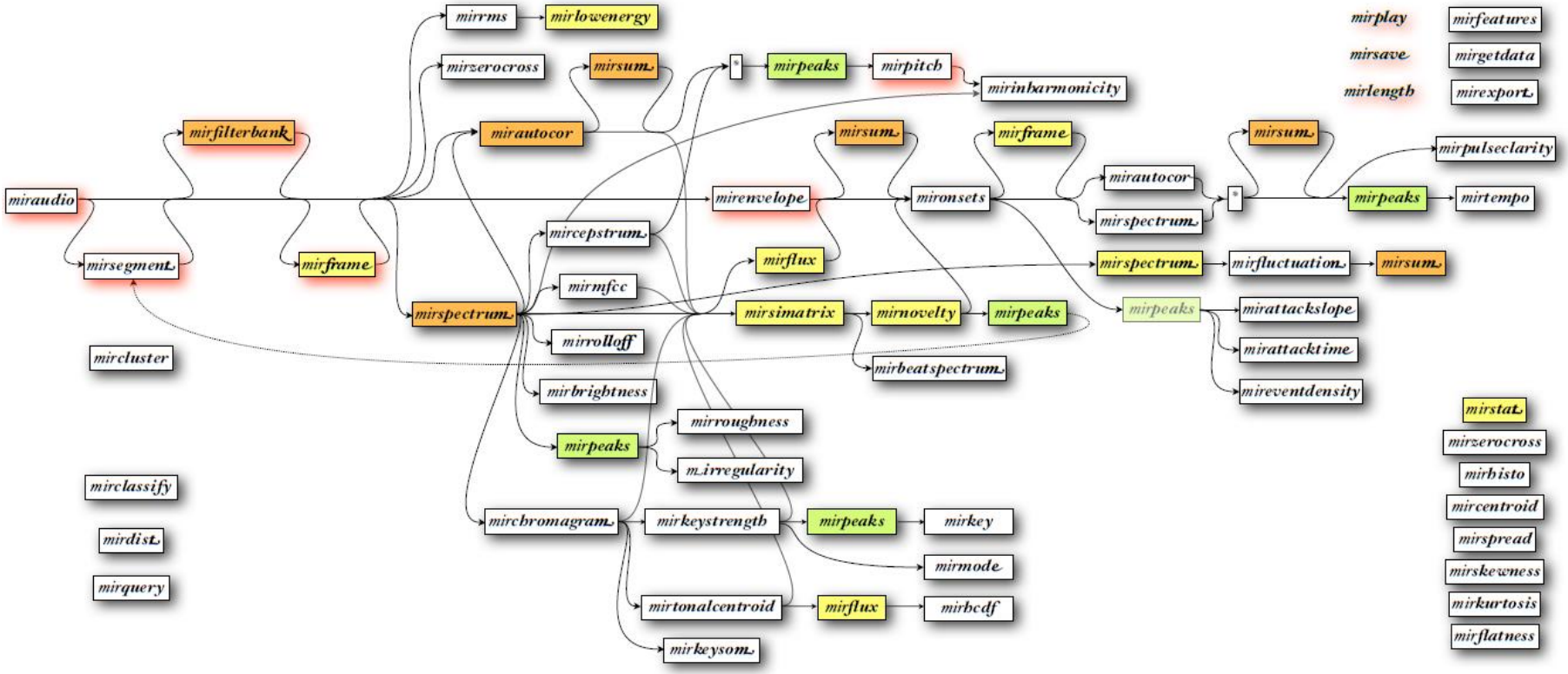


***Figura 1.*** *Esquema con las 50 funciones que forman MIRToolbox y su interrelación [20].*

## *2.3 Aprendizaje Automático*

El **aprendizaje automático** (en inglés, *machine learning*) persigue el desarrollo de algoritmos que otorguen a los computadores la capacidad de generalizar comportamientos e identificar patrones a partir de unos ejemplos previos dados en una fase de entrenamiento. En este sentido, el aprendizaje automático resulta adecuado para crear algoritmos que aprendan a identificar eventos salientes de cualquier tipo y, por ello, se ha aplicado en modelos acústicos existentes como los de G. E. Hinton y R. R. Salakhutdinov [21] y C. Zhang et al. [22] donde, más concretamente, se utilizan redes neuronales.

Además, el uso de redes neuronales profundas [23] se ha extendido en tareas como el procesado del habla y del audio ya que han producido buenos resultados. Su aplicación en tareas de clasificación de música es un buen ejemplo de ello [24]. Sin embargo, hay que tener en cuenta el alto requerimiento computacional de esta solución.

En los últimos años se han publicado, también, trabajos que utilizan modelos de regresión basados en redes neuronales [25], calculando un valor numérico para las variables de salida continuas en función de las entradas del sistema.

Para este estudio, se ha elegido la regresión como método para entrenar la red neuronal utilizada, como se explicará en el *capítulo 3.*

# Capítulo 3

## Modelo de la detección de la saliencia auditiva desarrollado

Para la creación de este modelo, teniendo en cuenta la dificultad de no contar con un desarrollo suficientemente amplio en este campo, el primer paso fue elegir una base de datos audiovisuales, ya que únicamente de audio no fue posible, lo suficientemente grande y representativa con la que poder extraer características de bajo nivel. En la *sección 3.1* se detallan sus características.

A continuación, se compararon varias funciones de la librería para MATLAB *MIRToolbox* para conseguir una buena caracterización de los archivos y que será comentada en la *sección 3.2*.

El siguiente paso, una vez elegidos el software sobre el que se iba a operar y los descriptores que mejor se adaptaban en este caso, fue elegir el tipo de red neuronal con el que entrenar el modelo desarrollado.

Finalmente, mediante la comparación con un valor umbral, se detectaron los eventos salientes de cada archivo.

Así, los pasos seguidos en el modelo desarrollado serían los siguientes:

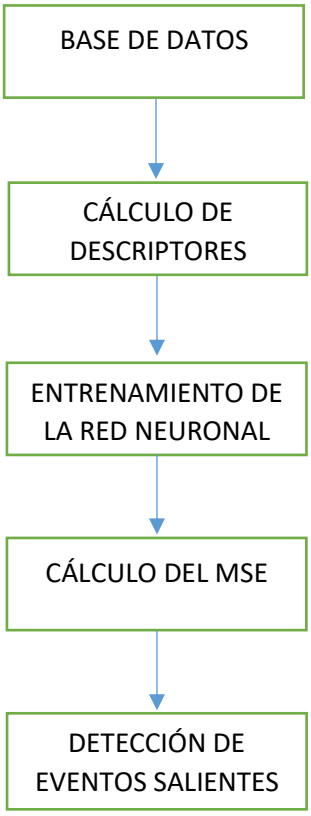


***Figura 2.*** *Esquema de los pasos del algoritmo desarrollado en el modelo.*

## *3.1. Base de datos*

Como ya se mostró en la *sección 2.1* del *Estado del arte* de este documento, existen varias bases de datos creadas en el contexto de la computación afectiva, sin embargo, la mayoría de ellas tienen alguna restricción que dificulta su utilización en ciertos proyectos en función del objetivo.

Por ejemplo, en la detección de saliencia auditiva, como es este caso, es imprescindible contar con archivos que contengan audio, éste fue el motivo por el que no se consideró la utilización de la base de datos *EMDB* [30] ya que no proporciona el audio de los vídeos que contiene.

Por otro lado, *DEAP* y *The Violent Dataset* [28, 31] se descartaron debido a que presentan problemas de *copyright.* La primera sólo facilita enlaces a la plataforma YouTube los cuales, en algunos casos, ya no están disponibles. En el caso de la segunda, el enlace es al sitio web de *Amazon* donde se pueden encontrar los DVD utilizados.

Respecto a las otras bases de datos mencionadas en el *Estado del arte*, *FilmStim* [27] contiene anotaciones globales, lo que no resulta adecuado a la hora de analizar emociones que duran, en ocasiones, unos pocos segundos.

Después de esta primera evaluación, quedaron tres bases, de las propuestas, como posibles candidatas para este proyecto: *HUMAINE, MAHNOB-HCI* y *LIRIS-ACCEDE* [26, 29, 1]. De ellas se eligió LIRIS-ACCEDE[2] por las siguientes razones:

- Las películas recopiladas pertenecen a diversos géneros audiovisuales, configurando una base de datos suficientemente representativa en cuanto a reproducción de todos los escenarios posibles [1].
- Todo su contenido se ha publicado con licencia *Creative Commons*, lo que facilita su acceso.

Más concretamente, LIRIS-ACCEDE está formada por 160 películas en total de las cuales 30 han pasado por un proceso de anotación continua, mediante un dispositivo de medición de la respuesta galvánica de la piel. Estas señales fisiológicas se midieron en un experimento en el que participaron 13 personas de nacionalidad francesa, de los cuales 11 eran mujeres y 2 hombres. Las edades de los sujetos estaban dentro del rango entre los 22 y los 45 años. Las anotaciones se tomaron durante 4 sesiones con las siguientes condiciones[3]:

- Cada sesión se realizó en un periodo del día distinto para minimizar la carga cognitiva de los participantes.
- Duraban aproximadamente 2 horas desde su inicio hasta la finalización de la toma de anotaciones.

[2] http://liris-accede.ec-lyon.fr/

[3] HAL Id: hal-01200730 https://hal.archives-ouvertes.fr/hal-01200730

- Los participantes firmaron un formulario otorgando su consentimiento.
- En cada sesión se informó del dispositivo de grabación de la respuesta galvánica de la piel, de que era totalmente seguro para la salud y de que se garantizaba el anonimato de los participantes.

El dispositivo electrónico mencionado se colocaba sobre los dedos de la mano. En la siguiente figura se muestra la apariencia de dicho dispositivo:

***Figura 3.*** *Dispositivo electrónico para la toma de anotaciones [18].*

Aunque el conjunto de datos fisiológicos recopilados en LIRIS-ACCEDE procede de 30 películas, para la realización de este experimento, se ha realizado una selección formada por 18 de ellas de manera que se redujese la carga computacional durante los experimentos a la vez que siguiera siendo representativa en cuanto a diversidad de género y duración. Las películas seleccionadas son las siguientes:

| ***Nombre*** | ***Duración*** | ***Género*** |
|---|---|---|
| After The Rain | 0 h 9 min 49 s | Drama |
| Barely Legal Stories | 0 h 16 min 28 s | Acción / comedia |
| Big Buck Bunny | 0 h 9 min 56 s | Animación |
| Cloudland | 0 h 11 min 41 s | Drama |
| Damaged Kung Fu | 0 h 16 min 54 s | Acción |
| First Bite | 0 h 10 min 40 s | Drama |
| Full Service | 0 h 18 min 41 s | Comedia |
| Islands | 0 h 2 min 53 s | Documental |
| Lesson Learned | 0 h 12 min 58 s | Documental / acción |
| Norm | 0 h 6 min 30 s | Comedia |
| Nuclear Family | 0 h 28 min 20 s | Drama |
| Riding The Rails | 0 h 15 min 0 s | Drama |
| Sintel | 0 h 14 min 48 s | Animación |
| Tears of Steel | 0 h 12 min 14 s | Ciencia ficción |
| The Room of Franz Kafka | 0 h 4 min 9 s | Alternativo |
| The Secret Number | 0 h 15 min 31 s | Suspense |
| To Claire from Sonny | 0 h 6 min 54 s | Romántico |
| You Again | 0 h 14 min 30 s | Romántico |

***Tabla 3.*** *Listado de películas que forman el conjunto de datos.*

El *ground truth*, entonces, se forma a partir de la medición de la *Respuesta Galvánica de la Piel (con sus siglas en inglés, GSR)* durante la visualización de estas películas. Llegados a este punto es importante considerar que, aunque este proyecto se centra en la detección de la saliencia auditiva, se ha utilizado una base de datos con contenido audiovisual, esto es así porque esta base proporciona fácil acceso y es representativa en cuanto a la diversidad de su contenido. Además, estas anotaciones se ven influidas por el audio de las películas en el momento del visionado con lo que, la extracción del audio de las películas se puede considerar como una buena aproximación a la realidad.

## *3.2. Elección de descriptores*

Para la extracción de características relacionadas con archivos de audio se eligió la herramienta de MATLAB *MIRToolbox*, ya que se disponía de experiencia programando en este lenguaje y la *Toolbox* está diseñada para el análisis de archivos musicales y de audio.

Como se explicó en el *capítulo 2.2* y, según se muestra en el gráfico de la *figura 1*, *MIRToolbox* contiene 50 funciones que se pueden aplicar al ámbito sonoro. De ellas se eligieron dos para el cálculo de los descriptores.

Esta elección se basó en la comparativa de varias de las funciones disponibles, valorando su adecuación al objetivo final de este proyecto. A continuación se exponen dichas funciones [20]:

- Mirspectrum

Esta función calcula la *Transformada Rápida de Fourier (con sus siglas en inglés, FFT)* mediante la función de MATLAB *fft* para obtener la descomposición de la energía de la señal repartida en amplitud en cada una de las frecuencias.

Esta función acepta modificaciones en el tamaño de ventana que utiliza *(frame)* y en el tipo de bandas de frecuencia, entre otros.

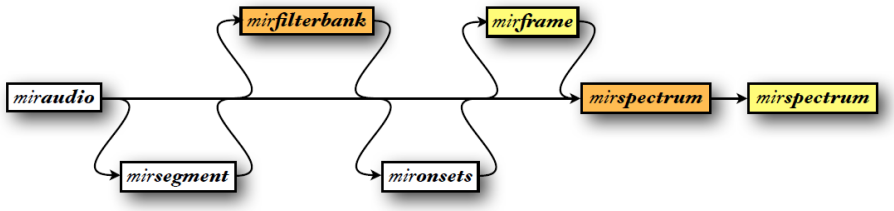


***Figura 4.*** *Diagrama de flujo de las interconexiones para mirspectrum [20].*

- Mirroughness

Realiza una estimación de la disonancia sensorial total mediante el procesado de los valores de pico del espectro de la señal. Después calcula la media entre cada pareja de picos posible.

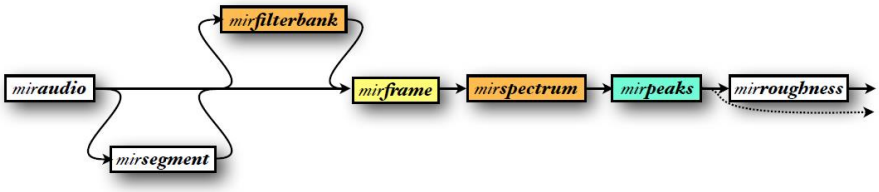


***Figura 5.*** *Diagrama de flujo de las interconexiones de mirroughness [20].*

- Mirattackslope

Esta función proporciona una estimación de la inclinación de la amplitud de la señal respecto del tiempo de ataque. En la siguiente figura se muestran mediante flechas estas medidas:

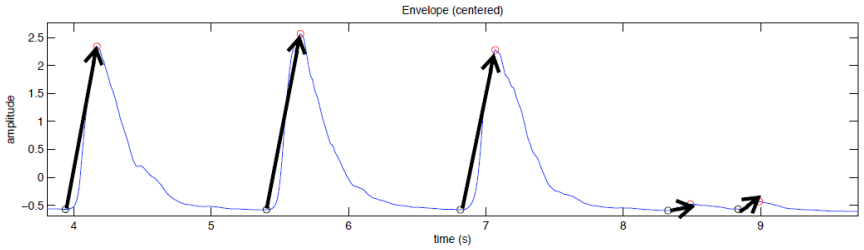


***Figura 6.*** *Ilustración del tiempo de ataque de una señal [20].*

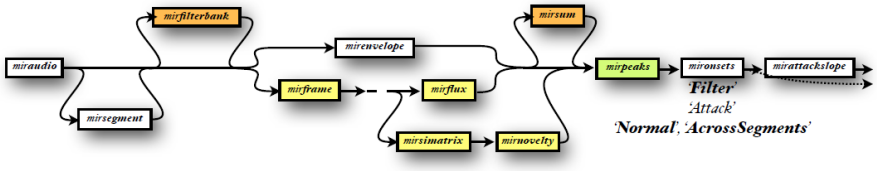


***Figura 7.*** *Diagrama de flujo de las interconexiones de mirattackslope [20].*

- Mirnovelty

La curva de novedad representa la probabilidad de que haya transiciones entre estados sucesivos a lo largo del tiempo, indicado por picos de señal, y su importancia relativa mediante las amplitudes de dichos picos.

Esta curva se calcula comparando, mediante correlación cruzada, los valores locales a lo largo de la diagonal de la matriz de similitud con un Kernel tipo “tablero de ajedrez” Gaussiano.

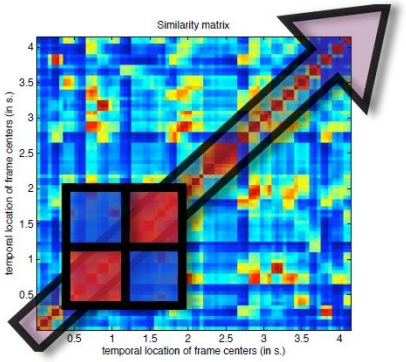


***Figura 8.*** *Ilustración de la matriz de similitud calculada para la función mirnovelty [20].*

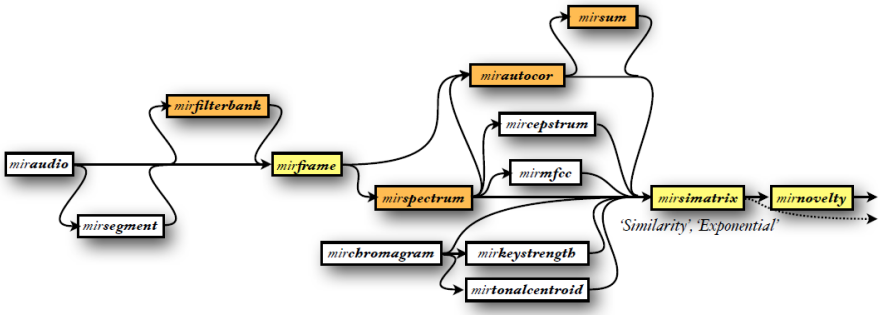


***Figura 9.*** *Diagrama de flujo de las interconexiones de mirnovelty [20].*

Finalmente se optó por utilizar las funciones *mirspectrum* y *mirnovelty* únicamente ya que la función *mirroughness* se centra en la detección de la disonancia sensorial lo que la hace demasiado específica en cuanto a que busca sonidos que producen, en general, rechazo en el oído y, por otro lado, *mirattackslope* presentaba el inconveniente de no aceptar entre sus parámetros un valor personalizado de *frame*, sin embargo, se había especificado en las demás el valor de *frame* a 10 con solapamiento del 50% y eso producía desajustes en el número de valores resultantes.

## *3.3. Red Neuronal*

Una vez obtenidos los valores proporcionados por los descriptores, se utilizan junto con los valores de *ground truth* procedentes de LIRIS-ACCEDE para entrenar la red neuronal mediante la *toolbox* de MATLAB: *Neural Network Fitting Tool*. Se escogió esta herramienta por su facilidad de uso y porque resultaba conveniente teniendo en cuenta que ya se había hecho uso de otra *toolbox* de MATLAB denominada *MIRToolbox*.

La interfaz gráfica de esta herramienta permite, de manera sencilla, seleccionar las entradas, los objetivos *(en inglés, targets)* y el número de neuronas según el criterio del usuario.

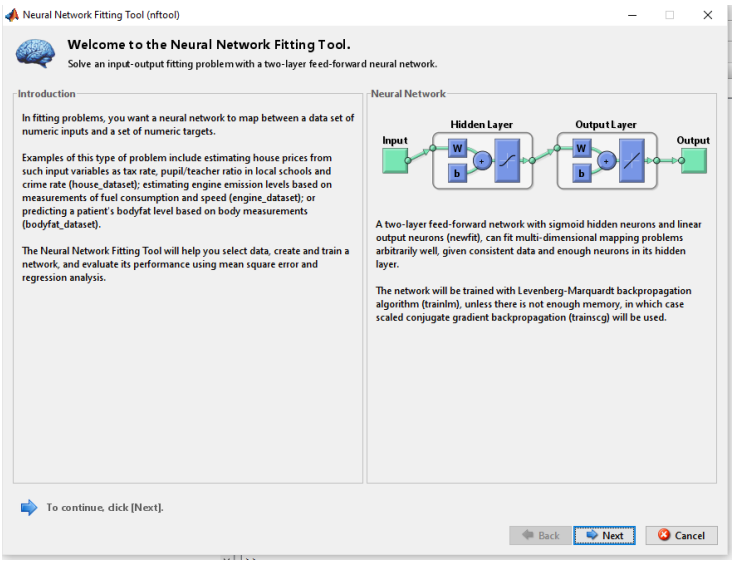


***Figura 10.*** *Interfaz gráfica de la herramienta Neural Network Fitting Tool de MATLAB [2].*

Para este modelo se ha entrenado una red neuronal por regresión y el MSE *(en inglés, Medium Squared Error)* para evaluar su comportamiento.

Según se informa en la pantalla de inicio de la interfaz, esta red estará formada por 2 capas hacia delante *(en inglés, feed-forward)* y serán entrenadas mediante un algoritmo de retro propagación *(en inglés, backpropagation).*

Por último, señalar que una de las posibilidades que ofrece esta herramienta es generar el código MATLAB correspondiente a cada operación realizada a través de las pantallas de la interfaz, opción que ha sido utilizada en este proyecto con el objetivo de automatizar el proceso para las 18 películas del conjunto seleccionado.

# Capítulo 4

## Experimentos y Resultados

Una vez definidas cada una de las partes del modelo creado para la detección de la saliencia auditiva, se procedió a su integración en un único sistema.

En este apartado van a especificarse los parámetros configurados para generar el modelo, obteniendo los valores de MSE como medida de la fiabilidad de los resultados de entrenar la red neuronal explicada en la *sección 3.3*.

### *4.1. Protocolo de Experimentación*

El protocolo seguido en este trabajo responde al siguiente orden: creación de la base de datos a partir de la de LIRIS-ACCEDE, extracción del audio de las películas seleccionadas, obtención de los descriptores, entrenamiento de la red neuronal, cálculo del MSE y de los parámetros indicadores del rendimiento del sistema. En el siguiente esquema se representan estos pasos con las herramientas utilizadas:

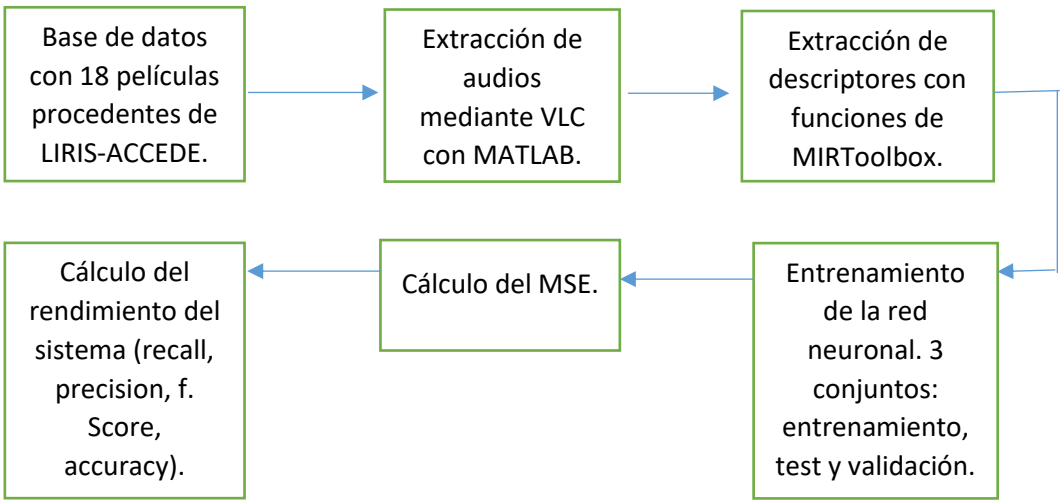


***Figura 11.*** *Esquema de los pasos seguidos durante el protocolo de experimentación.*

#### 4.1.1. Creación de la base de datos de audio

Como se ha explicado en la sección 3.1, LIRIS-ACCEDE está formada por 160 películas de las cuales 30 disponen también de anotaciones fisiológicas mediante la medida de la respuesta galvánica de la piel. También se ha comentado que de las 30, se eligieron 18, de manera aleatoria, para reducir el coste computacional.

Estas películas tienen formato *mp4* pero en este modelo se requieren archivos de audio en formato *wav*. Para obtenerlos, se utilizó el reproductor multimedia VLC de uso gratuito y código abierto mediante un comando proporcionado por la propia herramienta que se muestra en la siguiente figura:

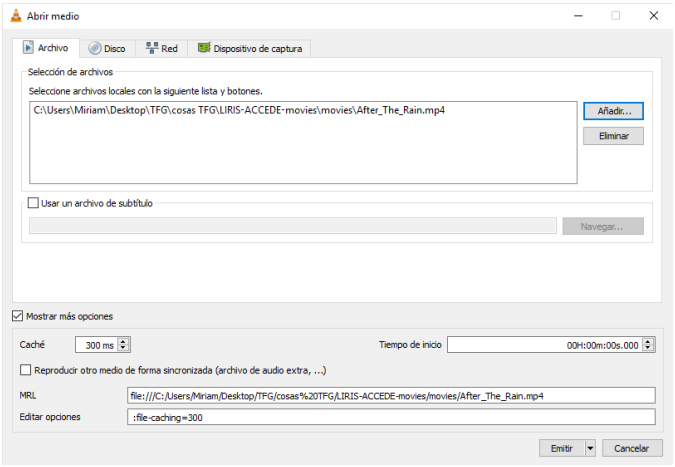


***Figura 12.*** *Interfaz para extraer el audio de un vídeo con VLC*[4]*.*

### 4.1.2. Descripción del ground truth

Tal y como se describió en el *apartado 3.1*, para generar el *ground truth* en este trabajo, se eligieron 18 películas de las 30 que tienen anotaciones fisiológicas de la base de datos LIRIS-ACCEDE.

Los valores de cada película se recogen en un archivo en formato *csv* en forma de matriz con tres columnas:

- La primera columna contiene el segundo de la película asociado a la medida fisiológica.
- Los valores de la *Respuesta Galvánica de la Piel (en inglés, GSR)* corresponden a la segunda columna.
- En la tercera, se muestran los valores de excitación recogidos mediante anotaciones de los participantes.

En este caso, mediante código MATLAB, se han utilizado los valores de GSR para conformar el *ground truth* empleado en el entrenamiento de la red neuronal que se detallará en el *capítulo 4.1.4.*

[4] http://www.videolan.org/vlc/

### 4.1.3. Configuración de los parámetros de los descriptores

El siguiente paso, después de separar el audio de las películas, es configurar los parámetros de las funciones empleadas para la extracción de los descriptores y así obtener una matriz de valores de tamaño: 41 para las filas, 40 proporcionados por la función mirspectrum y 1 por *mirnovelty* y el número de valores que tome la segunda columna del *ground truth*, con extensión *csv,* correspondiente a la medida de la GSR.

Las funciones de *MIRToolbox*, *mirspectrum* y *mirnovelty* se configuraron especificando un valor personalizado de *la longitud de trama* de 10 segundos y un solapamiento del 50%, estos valores se eligieron teniendo en cuenta que la frecuencia de muestreo del dispositivo para las medidas de la Respuesta Galvánica de la Piel era de 5 segundos.

Además, en la función *mirspectrum* se especificó la escala de las bandas de frecuencia con el parámetro *Mel,* quedando de la siguiente manera:

```
s = mirspectrum(nombre_wav, 'Frame', 10, 0.5, 'Mel');

n = mirnovelty(nombre_wav, 'Frame', 10, 0.5);
```

***Figura 13.*** *Funciones de MIRToolbox de MATLAB para el cálculo de los descriptores [20].*

### 4.1.4. Entrenamiento de la red neuronal

En el *capítulo 3.3*, se presentó la interfaz de la *Neural Network Fitting Tool* de MATLAB. En este apartado se van a definir los valores utilizados para el entrenamiento de la red neuronal y así obtener una medida de la calidad de los descriptores utilizados, mediante el cálculo del MSE.

Los parámetros configurables en esta herramienta y sus valores para este experimento son los siguientes:

- INPUTS: se refiere al archivo de entrada donde se proporcionan a la red los valores para el entrenamiento. En este caso, este fichero está en formato *.mat* y contiene los valores de los descriptores devueltos por las dos funciones usadas: *mirspectrum* y *mirnovelty.*
- TARGETS: contiene un archivo con los valores con los que se ajustará la red durante el periodo de entrenamiento. Para este trabajo, los valores se extraen, mediante código MATLAB, de la segunda columna del archivo *csv*, que son los que contienen el valor medido de la GSR.
- TRAINING: En este parámetro se asigna el porcentaje de muestras que se destinarán a entrenamiento de la red. Este parámetro toma el valor del 50% de las muestras seleccionadas aleatoriamente.

- VALIDATION: Es el parámetro donde se cuantifica el porcentaje de muestras que se utilizarán para el conjunto de validación. Toma un valor del 25%.
- TEST: En este parámetro se asigna el resto de porcentaje que quede después de asignar los dos conjuntos anteriores y se utiliza para evaluar la red neuronal después de entrenarla. En este caso, un 25%.
- HIDDEN NEURONS: es el número de neuronas que tendrá la capa oculta. En el experimento realizado se utilizaron los valores de 10, 15 y 20 neuronas para comparar.

Una vez entrenada la red para distinto número de neuronas, se obtienen los siguientes resultados para MSE que permiten elegir la configuración más óptima:

| *Nombre* | *MSE (10 neuronas)* | *MSE (15 neuronas)* | *MSE (20 neuronas)* |
|---|---|---|---|
| After The Rain | 0,00013 | 4,71e-05 | 0,00011 |
| Barely Legal Stories | 2,39e-05 | 2,76e-05 | 3,48e-05 |
| Big Buck Bunny | 7,01e-05 | 0,00012 | 8,76e-05 |
| Cloudland | 4,73e-05 | 3,77e-05 | 3,84e-05 |
| Damaged Kung Fu | 0,000105 | 0,000103 | 0,00016 |
| First Bite | 6,33e-05 | 0,00011 | 0,00018 |
| Full Service | 3,40e-05 | 7,67e-05 | 3,25e-05 |
| Islands | 0,0092 | 0,0044 | 0,0040 |
| Lesson Learned | 3,25e-05 | 0,00015 | 6,81e-05 |
| Norm | 0,00031 | 0,00038 | 0,00033 |
| Nuclear Family | 1,46e-05 | 1,22e-05 | 1,41e-05 |
| Riding The Rails | 5,83e-05 | 9,07e-05 | 6,85e-05 |
| Sintel | 3,48e-05 | 5,33e-05 | 3,87e-05 |
| Tears of Steel | 0,00018 | 0,00013 | 0,00014 |
| The Room of Franz Kafka | 0,0014 | 0,0016 | 0,0017 |
| The Secret Number | 2,40e-05 | 4,02e-05 | 3,15e-05 |
| To Claire from Sonny | 0,00046 | 0,00039 | 0,00083 |
| You Again | 9,45e-05 | 8,50e-05 | 4,20e-05 |

***Tabla 4.*** *Resultados obtenidos de MSE para cada película con distinto número de neuronas.*

En los experimentos llevados a cabo, se ha optado por ejecutar las películas de una sola vez, utilizando un código recursivo, a la hora de entrenar la red neuronal repartiendo un 50% de la misma para el conjunto de train, y un 25% para el de test y otro 25% para el de validación. Esto se ha configurado así para que en todos los conjuntos hubiese muestras representativas de cada película.

Como puede observarse en la tabla anterior, los resultados son similares en las distintas configuraciones de número de neuronas ocultas, no obstante, el experimento llevado a cabo con 10 neuronas produce el mínimo mse en 9 de los 18 archivos, mientras que en el caso de 15 neuronas el mínimo mse se da en 6 de los 18 y para 20 neuronas, en 5.

## *4.2. Resultados*

- Fórmula para el cálculo de la media:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Fórmula para el cálculo de la desviación típica:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

Media del MSE (con 10 neuronas) = $6.85 \cdot 10^{-4}$.

Media del MSE (con 15 neuronas) = $4.33 \cdot 10^{-4}$.

Media del MSE (con 20 neuronas) = $4.38 \cdot 10^{-4}$.

Desviación típica (con 10 neuronas) = $2.61 \cdot 10^{-2}$.

Desviación típica (con 15 neuronas) = $2.08 \cdot 10^{-2}$.

Desviación típica (con 20 neuronas) = $2.09 \cdot 10^{-2}$.

Una vez elegidos los parámetros con la configuración presentada en el apartado anterior y para un número de 15 neuronas en la capa oculta, se procede al cálculo de los parámetros asociados al rendimiento de la red neuronal que se muestran en el *apéndice B.*

Para cada película se han dibujado dos gráficas, la primera muestra el análisis de regresión del sistema y la segunda figura presenta la comparativa entre los valores de ground truth y los obtenidos por el sistema, ambos después de ser normalizados.

Estos parámetros proporcionan una visión global de la fiabilidad de la red neuronal entrenada. Cuando esta red alcanza un ajuste óptimo, los valores de MSE tienden a 0 y los correspondientes a la R de Pearson tienden a 1, esto significa que la diferencia entre los valores esperados y los obtenidos es casi nula, para el caso del MSE, y que la correlación entre estos valores es alta.

Si se observan los valores de cada gráfica, se llega a la conclusión de que dichos valores no son óptimos a nivel global, ya que en algunos casos, como el de la película *The Room of Franz Kafka* o *Cloudland,* el valor de la R de Pearson queda muy por debajo del valor ideal pero, teniendo en cuenta que los datos de *ground truth* utilizados han sido extraídos de contenido audiovisual y, sin embargo, en este trabajo, se han utilizado únicamente descriptores para audio, los resultados se ajustan bastante bien a lo que podría esperarse.

Otra de las razones por las que el modelo podría no generalizar bien es porque algunos archivos son demasiado cortos en el tiempo o en otros casos el audio extraído es insuficiente debido al tipo de documento como es el caso de *To Clair from Sony* en el que, en toda la película el audio

asociado es una banda sonora de fondo con una pequeña narración con lo que resulta bastante monótono y además su duración es de seis minutos aproximadamente.

Habiendo obtenido unos valores de salida mediante la red neuronal entrenada, procedemos al cálculo de los sucesos salientes de cada archivo de audio. Para ello, primero fijamos un umbral con valor igual a la media de los valores de referencia de cada película y después para ajustar mejor dicho umbral usamos la media de los valores de referencia más la desviación típica de la salida, es decir, la media de los valores obtenidos mediante la medición de la GSR más la desviación típica calculada para la salida. En este sentido, todos los valores obtenidos mediante este algoritmo que sobrepasen dicho umbral se considerarán eventos salientes.

Para conocer el rendimiento del sistema en la detección explicada en el párrafo anterior, se han calculado los siguientes parámetros:

| Nombre | Precision | Recall | F. Score |
|---|---|---|---|
| After The Rain | 0,48 | 0,72 | 0,58 |
| Barely Legal Stories | 0,57 | 0,46 | 0,51 |
| Big Buck Bunny | 0,53 | 0,50 | 0,51 |
| Cloudland | 0,54 | 0,55 | 0,54 |
| Damaged Kung Fu | 0,55 | 0,44 | 0,49 |
| First Bite | 0,55 | 0,45 | 0,49 |
| Full Service | 0,56 | 0,40 | 0,47 |
| Islands | 0,55 | 0,41 | 0,47 |
| Lesson Learned | 0,58 | 0,43 | 0,50 |
| Norm | 0,58 | 0,43 | 0,49 |
| Nuclear Family | 0,58 | 0,35 | 0,44 |
| Riding The Rails | 0,58 | 0,35 | 0,43 |
| Sintel | 0,59 | 0,35 | 0,44 |
| Tears of Steel | 0,59 | 0,33 | 0,43 |
| The Room of Franz Kafka | 0,59 | 0,34 | 0,43 |
| The Secret Number | 0,57 | 0,37 | 0,45 |
| To Claire from Sonny | 0,57 | 0,38 | 0,46 |
| You Again | 0,57 | 0,37 | 0,45 |

***Tabla 5.*** *Valores de Precision, recall y F. Score para cada película con umbral igual a la media de los valores de ground truth, sin normalización.*

| Nombre | Precision | Recall | F. Score |
|---|---|---|---|
| After The Rain | 0.60 | 0.39 | 0.47 |
| Barely Legal Stories | 0.41 | 0.29 | 0.34 |
| Big Buck Bunny | 0.53 | 0.48 | 0.50 |
| Cloudland | 0.35 | 0.24 | 0.28 |
| Damaged Kung Fu | 0.44 | 0.61 | 0.51 |
| First Bite | 0.73 | 0.30 | 0.42 |
| Full Service | 0.24 | 0.74 | 0.36 |
| Islands | 0.75 | 0.75 | 0.75 |
| Lesson Learned | 0.65 | 0.52 | 0.58 |
| Norm | 0.89 | 0.50 | 0.64 |

| | | | |
|---|---|---|---|
| Nuclear Family | 0.53 | 0.39 | 0.45 |
| Riding The Rails | 0.68 | 0.63 | 0.65 |
| Sintel | 0.56 | 0.58 | 0.57 |
| Tears of Steel | 0.61 | 0.53 | 0.57 |
| The Room of Franz Kafka | 0.20 | 0.28 | 0.23 |
| The Secret Number | 0.22 | 0.24 | 0.23 |
| To Claire from Sonny | 0.20 | 0.60 | 0.30 |
| You Again | 0.57 | 0.53 | 0.55 |

***Tabla 6.*** *Valores de Precision, recall y F. Score para cada película con umbral igual a la media de los valores de ground truth más la desviación típica de la salida, con normalización.*

# Capítulo 5

# Conclusions and Future Work

## *5.1. Conclusions*

There are very few and recent studies about aural saliency detection. In this Project, we wanted to develop a model that Works with a large and representative database to obtain the targets and the measurements from descriptors with the objective of train a neural network by regression method.

However, if we analyze MSE and standard deviation values in some files, the results are not the best ones because some films of the database are more salient in image characteristics than the audio. For example, this is the case of *Islands movie*, this film is monotonous in the audio but is posible to obtain salient events in the image processing.

Despite this, we can say that the descriptors used in this Project are good enough to obtain better results with a database more focused in audio events.

## *5.2. Future work*

In the state of the art of this Project, we explained that aural saliency detection is a very recent model to obtain a good performance of event classification tools. For this reason, is a field where could be possible to create and improve research models.

Firstly, the database could be improved taking into consideration that some selected films could have few information in the audio and it can be a problem for the training process.

Secondly, more descriptors could be added making a more complex algorithm to use statistical moments like temporal and spectral centroids, apart from the energy of the signal and novelty curve.

Finally, as we mentioned in this paper, a better model can be performed joining a visual saliency model with this one to get a multimodal work for detection.

# Capítulo 6

# Presupuesto

Este capítulo se divide en dos apartados, uno para el detalle de la inversión en tiempo de la realización de este Trabajo Fin de Grado y el otro para la parte de la inversión económica.

## *6.1 Gestión del tiempo*

En esta sección se especifican las tareas realizadas desde el inicio hasta el final del proceso de desarrollo del proyecto.

En la siguiente tabla se muestran las tareas principales que se enumeraron en el capítulo 1, sección 1.3, relacionadas con el tiempo invertido en cada una de ellas:

| FASES DEL PROYECTO | ACTIVIDADES | HORAS INVERTIDAS |
|---|---|---|
| *Investigación y documentación* | Investigación de proyectos existentes | 45 horas |
| | Búsqueda de descriptores adecuados | 24 horas |
| *Desarrollo de software* | Elección de librerías adecuadas | 30 horas |
| | Creación de código MATLAB | 130 horas |
| *Ejecución del código* | Ajuste de parámetros | 45 horas |
| *Análisis de resultados y memoria* | Análisis de resultados | 24 horas |
| | Redacción de la memoria | 85 horas |
| *Total* | - | *383 horas* |

***Tabla 7.*** *Fases del proyecto, divididas en actividades con la asociación de su duración en horas.*

A esta relación deben añadirse las tutorías acordadas con la tutora de este proyecto, Dª Carmen Peláez Moreno.

## *6.2 Costes Asociados*

Para el cálculo de los costes asociados, se utilizarán los resultados del apartado anterior, referidos al tiempo y se considerará el sueldo medio para un profesor titular y un estudiante, considerado como auxiliar en las tablas de retribución salarial de la Federación Estatal Sectorial

de la Unión General de Trabajadores (FETE-UGT)[5]. Además, se considerará el coste material separándose en coste para el Hardware y coste asociado al Software.

| HARDWARE | | | | | | |
|---|---|---|---|---|---|---|
| *Herramienta* | *Características* | Coste (sin IVA) | Dedicación | Periodo de depreciación | Porcentaje de uso | Coste imputable |
| *Ordenador portátil* | Acer Aspire E1-571G<br>*Intel core i7-3612QM 2.1 GHz*<br>*500 GB HDD* | 640 € | 3 meses | 60 meses | 100% | **32 €** |

***Tabla 8.*** *Costes asociados al hardware utilizado en el proyecto.*

| SOFTWARE | |
|---|---|
| *Descripción* | *Valor económico* |
| *Licencia MATLAB (uso académico)* | 500 € |
| *Neural Network Toolbox (uso académico)* | 200 € |
| *Librería MIRToolbox* | Licencia gratuita |
| *VLC* | Licencia gratuita |
| *TOTAL* | *700 €* |

***Tabla 9.*** *Costes asociados al software utilizado en el proyecto.*

| COSTES DE PERSONAL | | | |
|---|---|---|---|
| *Cargo* | *Número de horas dedicadas* | *Coste / hora* | *Coste Total* |
| *Estudiante (auxiliar)* | 383 horas | 18,7 € / hora | 7.162,1 € |
| *Profesor Titular* | 110 horas | 32,31 € / hora | 3.554,1 € |
| *Total* | | | *10.716,2 €* |

***Tabla 10.*** *Costes asociados al personal para este proyecto.*

[5] Tabla salarial para el año 2015: http://www.feteugt.es/Data/UPLOAD/PRI-tablas salariales-XIII-convenio-centros-educacion-universitaria-2015.pdf

| COSTE TOTAL | |
|---|---|
| *Coste del equipamiento hardware* | 32 € |
| *Coste del equipamiento software* | 700 € |
| *Coste de personal* | 10.716,2 € |
| *IVA (21%)* | 2.404,12 € |
| *TOTAL* | *13.852,32 €* |

***Tabla 11.*** *Coste total para este proyecto.*

## Abstract

### *1. Introduction*

Saliency is the concept that defines the human brain's reaction which prioritizes certain stimuli over other less important to know the world around us. In this context, aural salience is referred to auditory stimuli.

Recently, an increasing interest from the reasearch community and new development in measurement systems have allowed the publication of studies such as [6] by T. Tsuchida y G. W. Cottrell or [17] by F. Tordini et al., also, specific databases have been created for this tasks such as HUMAINE [26] or LIRIS-ACCEDE [1] which is used in this work.

The objective of this project is to select descriptors to create a model to obtain good results when an artificial neural network is trained. For this purpose, it is used measurements from Galvanic Skin Response (GSR) as ground truth.

The firs step in this process was research and documentation period, after we knew the recently published documents, we could start to develop and execute the software for this project in order to get the input of the neural network and later, we took a decision about event detection with the result. Finally, this paper was written with this results.

The socioeconomic environment of this work is defined by the applications in advertising media, and event detection.

There are no specific laws for the field of saliency detection due to the difficulties of gathering all applications in the same context. However, in Spain, the data protection law [BIBLIO] applies to it.

### *2. State of the Art*

In saliency research there are two main concepts, bottom-up salience and top-down detection. Then, the difference between them is explained:

- Bottom-up saliency is stimuli oriented, attention is drawn involuntarily.
- Top-down detection is a slower process and the subject is consciously focused on the event.

First studies related to aural salience were based on bottom-up model and were adapted from the visual saliency works [3]. This is the reason why the characteristics used in them were not representative enough.

A new configuration to improve bottom-up saliency models is emerging working with bio-inspired low-level features because is the best way to imitate the human auditory system. However, about top-down detection, the state of development is lower.

When it comes a more complex model than those mentioned, there is another way to detect saliency with both models combined, that is, *bottom-up* and *top-down* working together [5] [16].

In particular, for this project we chose the bottom-up model considering that the stimuli came from outside, in which, the subject doesn´t have control over it.

One of the most difficult task in this work was to find reliable ground truth data because it is not easy to meassure, objectively, data about the reaction of our sense of hearing.

For this reason, different alternatives have been studied in this field:

- The subject decide if the stimuli is salient subjectively [4] [6].
- The same as the previous but adding some objective parameters like the response time [17].
- Manual audio tagging through an interface.

Several databases according to this procedures have been published during this years to obtain a reliable ground truth. However, they are still, in general, small in size and they have copyright problems.

Followed, there are some of the latest:

| ***Name*** | ***Size*** |
|---|---|
| HUMAINE | *50 clips between 5 seconds and 3 minutes.* |
| FilmStim | 70 extracts of films between 1 minute y 7 minutes. |
| DEAP | 120 musical videos 1 minute. |
| MAHNOB-HCI | 20 extracts of films between 35 y 117 seconds. |
| EMDB | 52 clips of 40 seconds without audio. |
| VIOLENT SCENES DATASET | 25 films. |
| LIRIS-ACCEDE | 9800 extracts of 160 films between 8 y 12 seconds. |

***Table 12.*** *Diferent databases published to detect emotions [1].*

This databases are very specific so they are not the best option in some cases. They have not been created as a standard database.

On the other hand, it is important to choose the suitable descriptors in order to obtain a model that let us to detect salient stimuli properly. As we said, there is not many researches for auditory salience, but the few that exist are based on:

- Waveform envelope

- Tone
- Bandwidth
- Parameters related to Gammatone filterbank analysis
- Parameters based on temporal and spectral centroids

In addition, there are some libraries for the analysis of musical files like the *MATLAB Toolbox* called *MIRToolbox* that perform this analysis through some functions that calculate different parameters according to our needs.

***Machine Learning:***

Machine learning is the expression used to speak of algorithm develop that allow computers to identify human behaviour and generalize it through previous examples given in a training phase.

For this reason, it is useful to detect salient events through trained algorithms and used with neural networks.

In recent years, some research have been published using regression models based on neural networks [25]. This research calculate numerical values for the output depending on the inputs.

In our case, we used a regression model to train a neural network as we mentioned previously.

## *3. Model development*

To create this model, and considering that this field is not developed enough, we chose an audiovisual database big enough to extract different low level characteristics.

After that, we compared some functions of MIRToolbox library to obtain a good characterization of the files and, finally, we trained the neural network with the data and we detected the salient event by comparing the output with a threshold.

### *3.1. Database*

In our research, we needed a database with audio files, so we had to discard the EMDB database [30]. Furtheremore, DEAP and The Violent Dataset [28, 31] were discarded because they have copyright restrictions and it is difficult, in some cases, to access some files.

The rest of the databases were not chosen because they are not the best option to analyze emotions.

Finally, we chose LIRIS-ACCEDE for the following reasons:

- Films in this database have different genres, so they are more representative than others databases [1].
- They are published under *Creative Commons license*.

More specifically, LIRIS-ACCEDE is formed by 160 films within 30 have been processed with a *Galvanic Skin Response* device. This device was used in an experiment to obtain contiuous annotations.

In this experiment, we used 18 films to reduce the computational load. The selected movies are:

| ***Name*** | ***Duration*** | ***Genre*** |
|---|---|---|
| After The Rain | 0 h 9 min 49 s | Drama |
| Barely Legal Stories | 0 h 16 min 28 s | Action / comedy |
| Big Buck Bunny | 0 h 9 min 56 s | Cartoon |
| Cloudland | 0 h 11 min 41 s | Drama |
| Damaged Kung Fu | 0 h 16 min 54 s | Action |
| First Bite | 0 h 10 min 40 s | Drama |
| Full Service | 0 h 18 min 41 s | Comedy |
| Islands | 0 h 2 min 53 s | Documental |
| Lesson Learned | 0 h 12 min 58 s | Documental / acction |
| Norm | 0 h 6 min 30 s | Comedy |
| Nuclear Family | 0 h 28 min 20 s | Drama |
| Riding The Rails | 0 h 15 min 0 s | Drama |
| Sintel | 0 h 14 min 48 s | Cartoon |
| Tears of Steel | 0 h 12 min 14 s | Science fiction |
| The Room of Franz Kafka | 0 h 4 min 9 s | Alternative |
| The Secret Number | 0 h 15 min 31 s | Suspense |
| To Claire from Sonny | 0 h 6 min 54 s | Romantic |
| You Again | 0 h 14 min 30 s | Romantic |

***Table 13.*** *Films chose for the experiment.*

## *3.2. Descriptors*

To obtain the results we chose two functions of MIRToolbox that have good characteristics for our purpose:

- Mirspectrum

This function calculate FFT to decompose the signal energy on frecuency bands.

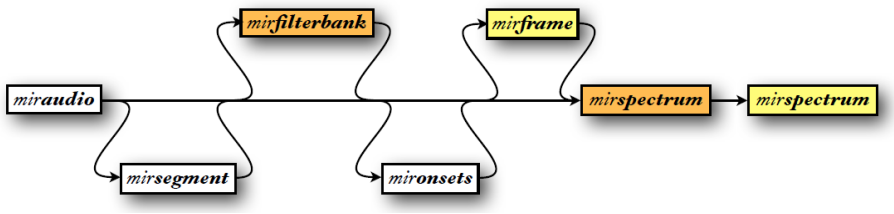


***Figure 14.*** *Mirspectrum [20].*

- Mirnovelty

This function calculates the novelty curve based on the probability of transitions between successive values in time.

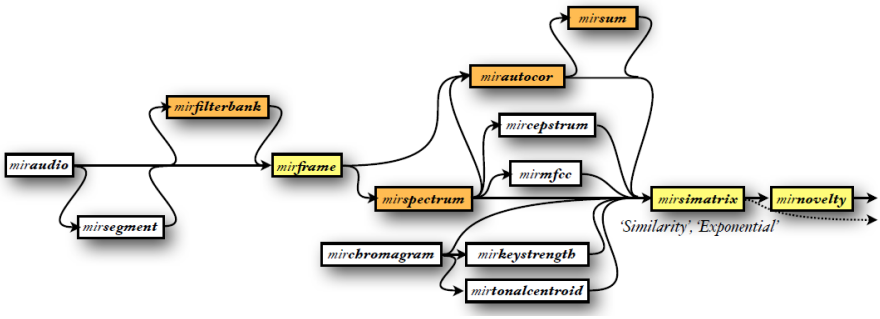


***Figure 15.*** *Mirnovelty [20].*

We used *mirnovelty* and *mirspectrum* with a frame value of 10 and 50% overlap.

## *3.3. Neural Network*

We used in this research the Neural Network Fitting Tool of MATLAB because this software was used with MIRToolbox and this is a good tool to train a network, it is, the graphic interface is easy to use and it has many configurable parameters.

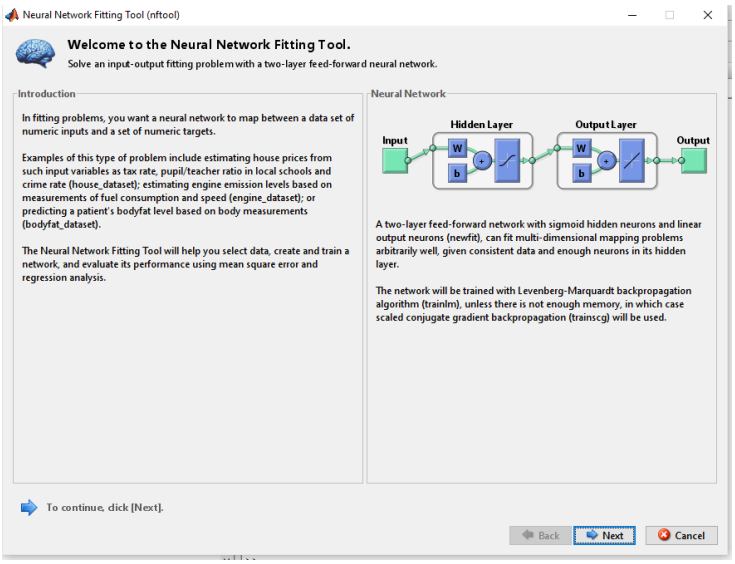


***Figure 16.*** *Neural Network Fitting Tool graphical interface of MATLAB [2].*

Specifically, we trained a neural network by regression with two feed-forward layers and backpropagation algorithm.

### *4. Experiments and results*

The next figure shows the experiment protocol followed in our work:

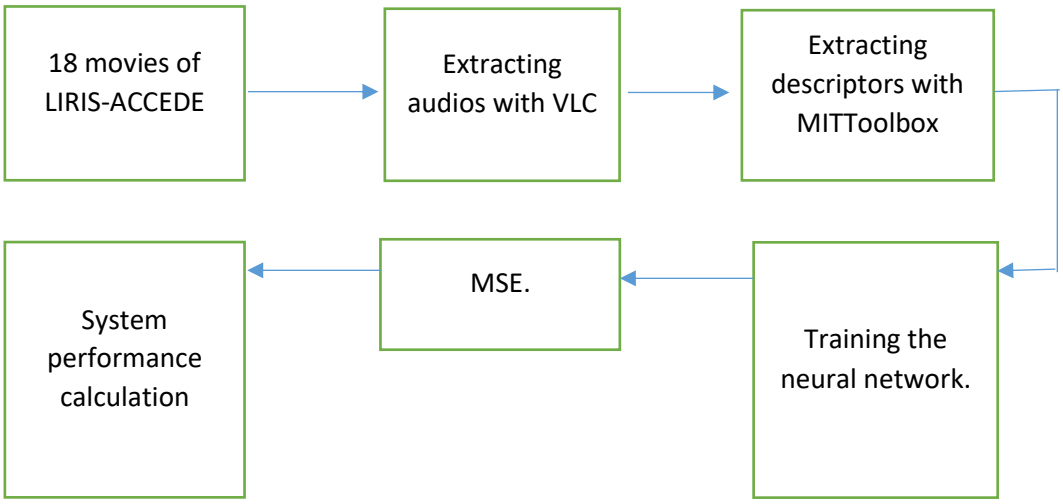


***Figure 17.*** *Scheme of the protocol.*

When we started the research, we chose 18 movies of LIRIS-ACCEDE database and we extracted the audio part of the movies with VLC[5].

The next step was to configure the descriptors parameters, as we mentioned we chose 10 sconds of frame length and 50% overlap to be able to compare with *Galvanic Response* values of the target set.

Later, we trained the neural network that is formed by:

- INPUTS: values obtained by mirspectrum and mirnovelty functions.
- TARGETS: composed by *galvanic skin response* values.
- TRAINING: 50% of the input values.
- VALIDATION: 25% of the input values.
- TEST: 25% of the input values.
- HIDDEN NEURONS: neurons of the hidden layer. In this case: 10, 15 and 20 to compair between them.

When we made the training part, at first, we obtained the next values for MSE:

| ***Name*** | ***MSE (10 neurons)*** | ***MSE (15 neurons)*** | ***MSE (20 neurons)*** |
|---|---|---|---|
| After The Rain | 0,00013 | 4,71e-05 | 0,00011 |
| Barely Legal Stories | 2,39e-05 | 2,76e-05 | 3,48e-05 |

| | | | |
|---|---|---|---|
| Big Buck Bunny | 7,01e-05 | 0,00012 | 8,76e-05 |
| Cloudland | 4,73e-05 | 3,77e-05 | 3,84e-05 |
| Damaged Kung Fu | 0,000105 | 0,000103 | 0,00016 |
| First Bite | 6,33e-05 | 0,00011 | 0,00018 |
| Full Service | 3,40e-05 | 7,67e-05 | 3,25e-05 |
| Islands | 0,0092 | 0,0044 | 0,0040 |
| Lesson Learned | 3,25e-05 | 0,00015 | 6,81e-05 |
| Norm | 0,00031 | 0,00038 | 0,00033 |
| Nuclear Family | 1,46e-05 | 1,22e-05 | 1,41e-05 |
| Riding The Rails | 5,83e-05 | 9,07e-05 | 6,85e-05 |
| Sintel | 3,48e-05 | 5,33e-05 | 3,87e-05 |
| Tears of Steel | 0,00018 | 0,00013 | 0,00014 |
| The Room of Franz Kafka | 0,0014 | 0,0016 | 0,0017 |
| The Secret Number | 2,40e-05 | 4,02e-05 | 3,15e-05 |
| To Claire from Sonny | 0,00046 | 0,00039 | 0,00083 |
| You Again | 9,45e-05 | 8,50e-05 | 4,20e-05 |

***Table 14.*** *MSE results for every film and different number of neurons.*

Comparing the values obtained, we chose 15 neurons to calculate the rest of the parameters. The graphical results for this number of neurons are presented in *Apendix B.*

The first figure shows the regression analysis and the second figure shows the comparison between ground truth values and output values after normalization process.

If we analyse the results, we can see that some of them are not optimum, this is because the ground truth used here have been extracted by a film with audio and video but we use only audio so it is not the same. Furthermore, some of this files are short in time and the algortihm doesn't have time enough to learn and works properly, an example of this is the film *Islands.*

After the train process we can detect the salient events using a threshold value calculated by the average of target set and the std of the outputs. This is posible comparing the output values with this threshold and calculating some parameters that show us if it works properly.

| Name | Precision | Recall | F. Score |
|---|---|---|---|
| After The Rain | 0.60 | 0.39 | 0.47 |
| Barely Legal Stories | 0.41 | 0.29 | 0.34 |
| Big Buck Bunny | 0.53 | 0.48 | 0.50 |
| Cloudland | 0.35 | 0.24 | 0.28 |
| Damaged Kung Fu | 0.44 | 0.61 | 0.51 |
| First Bite | 0.73 | 0.30 | 0.42 |
| Full Service | 0.24 | 0.74 | 0.36 |
| Islands | 0.75 | 0.75 | 0.75 |
| Lesson Learned | 0.65 | 0.52 | 0.58 |
| Norm | 0.89 | 0.50 | 0.64 |
| Nuclear Family | 0.53 | 0.39 | 0.45 |
| Riding The Rails | 0.68 | 0.63 | 0.65 |
| Sintel | 0.56 | 0.58 | 0.57 |
| Tears of Steel | 0.61 | 0.53 | 0.57 |

| | | | |
|---|---|---|---|
| The Room of Franz Kafka | 0.20 | 0.28 | 0.23 |
| The Secret Number | 0.22 | 0.24 | 0.23 |
| To Claire from Sonny | 0.20 | 0.60 | 0.30 |
| You Again | 0.57 | 0.53 | 0.55 |

***Table 15.*** *Precision, recall and F. Score of every film using normalized values.*

## *5.1. Conclusions*

There are very few and recent studies about aural saliency detection. In this Project, we wanted to develop a model that Works with a large and representative database to obtain the targets and the measurements from descriptors with the objective of train a neural network by regression method.

However, if we analyze MSE and standard deviation values in some files, the results are not the best ones because some films of the database are more salient in image characteristics than the audio. For example, this is the case of *Islands movie*, this film is monotonous in the audio but is posible to obtain salient events in the image processing.

Despite this, we can say that the descriptors used in this Project are good enough to obtain better results with a database more focused in audio events.

## *5.2. Future work*

In the state of the art of this Project, we explained that aural saliency detection is a very recent model to obtain a good performance of event classification tools. For this reason, is a field where could be possible to create and improve research models.

Firstly, the database could be improved taking into consideration that some selected films could have few information in the audio and it can be a problem for the training process.

Secondly, more descriptors could be added making a more complex algorithm to use statistical moments like temporal and spectral centroids, apart from the energy of the signal and novelty curve.

Finally, as we mentioned in this paper, a better model can be performed joining a visual saliency model with this one to get a multimodal work for detection.

# *6. Budget*

The next table shows the time invested in this research:

| PHASES | HOURS INVESTED |
|---|---|
| *Research and documentation* | 45 H. |
| | 24 H. |
| *Software development* | 30 H. |
| | 130 H. |
| *Code execution* | 45 H. |
| *Analisys of results* | 24 H. |
| | 85 H. |
| *Total* | *383 H.* |

***Table 16.*** *Time invested.*

To this relationship should be added the tutorials agreed with the tutor of this project, Mrs. Carmen Peláez Moreno.

The next table shows the total cost of this project:

| TOTAL COST | |
|---|---|
| *Hardware* | 32 € |
| *Software* | 700 € |
| *Staff resources* | 10.716,2 € |
| *IVA (21%)* | 2.404,12 € |
| *TOTAL* | *13.852,32 €* |

***Table 17.*** *Total cost for this project.*

## Medida de la calidad de la Red Neuronal

### 1. *Gráficas para el valor MSE y la R de Pearson de cada archivo*

- After The Rain

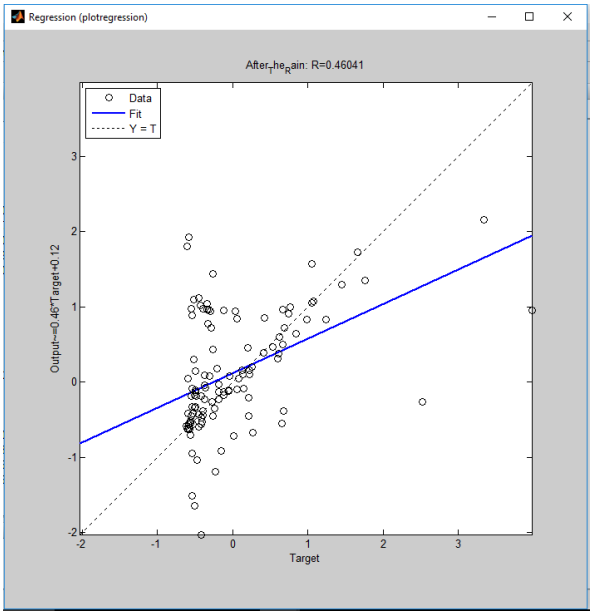


***Figura 18.*** *Regresión lineal de la película After The Rain.*

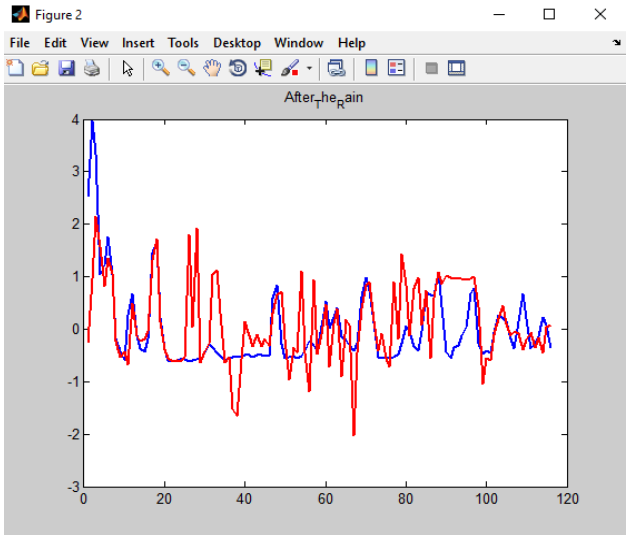


***Figura 19.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Barely Legal Stories

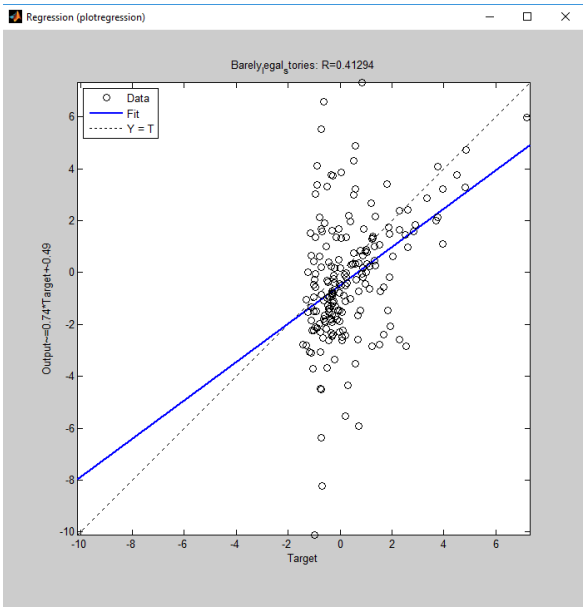


***Figura 20.*** *Regresión lineal de la película Barely legal stories.*

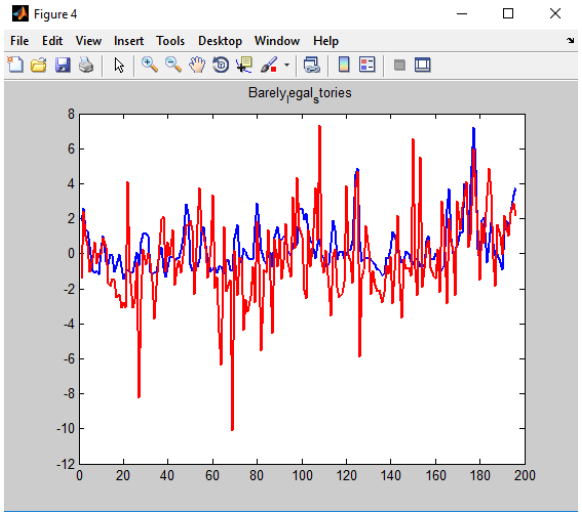


***Figura 21.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Big Buck Bunny

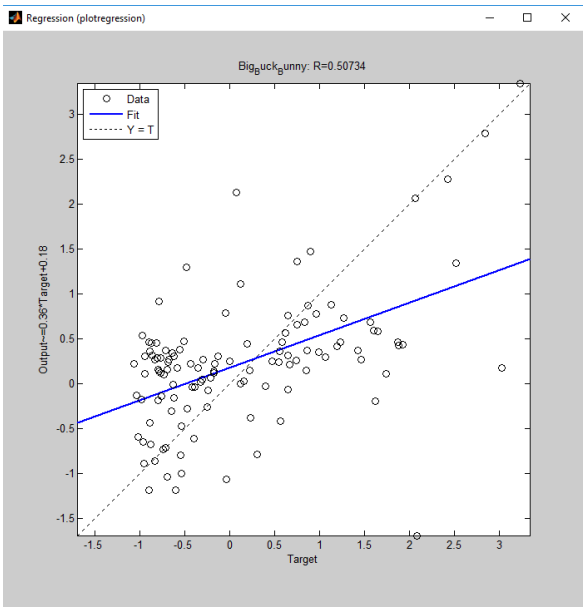


***Figura 22.*** *Regresión lineal de la película Big Buck Bunny.*

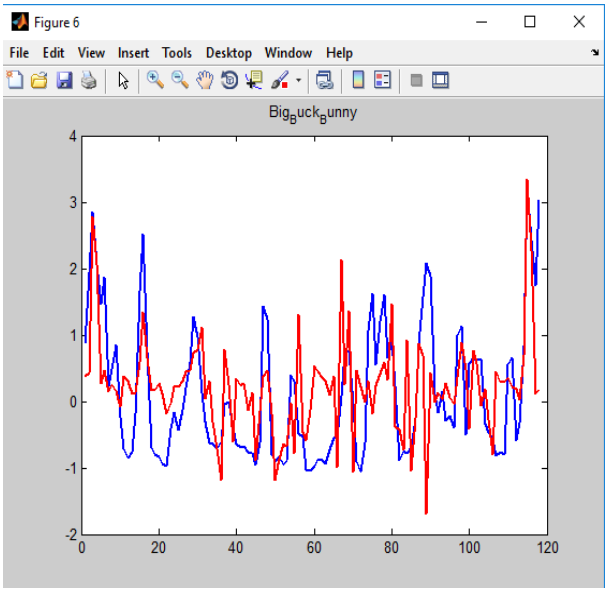


***Figura 23.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Cloudland

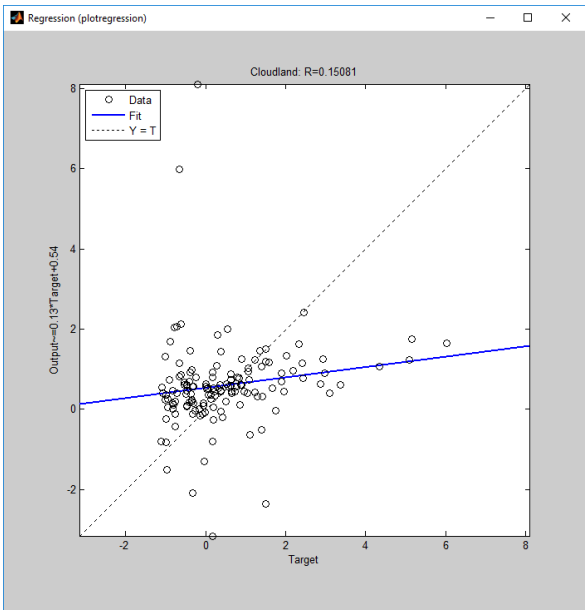


***Figura 24.*** *Regresión lineal de la película Cloudland.*

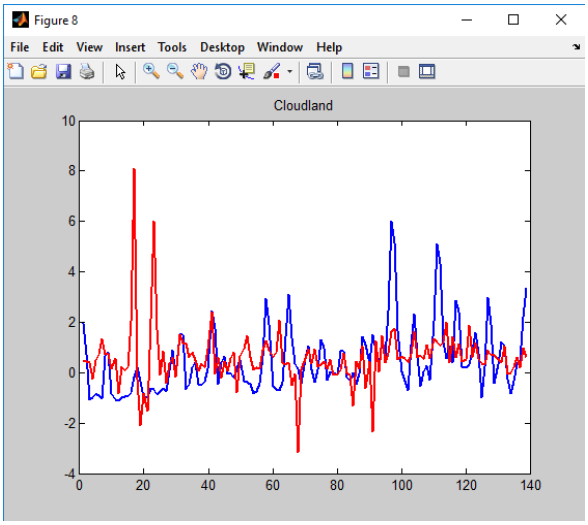


***Figura 25.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Damaged Kung Fu

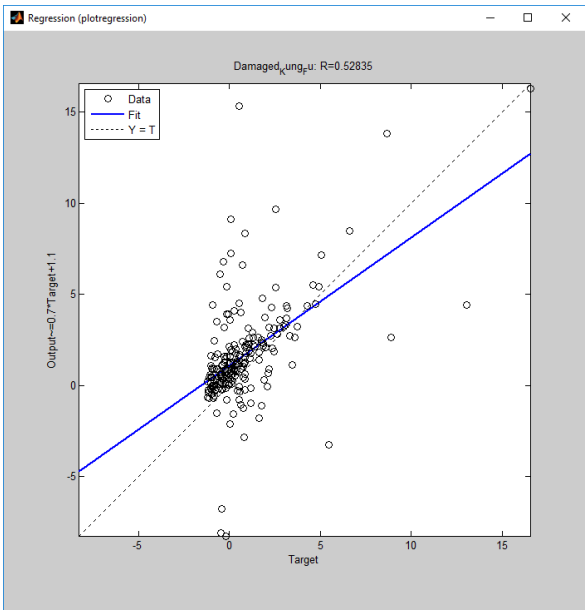


***Figura 26.*** *Regresión lineal de la película Damaged Kung Fu.*

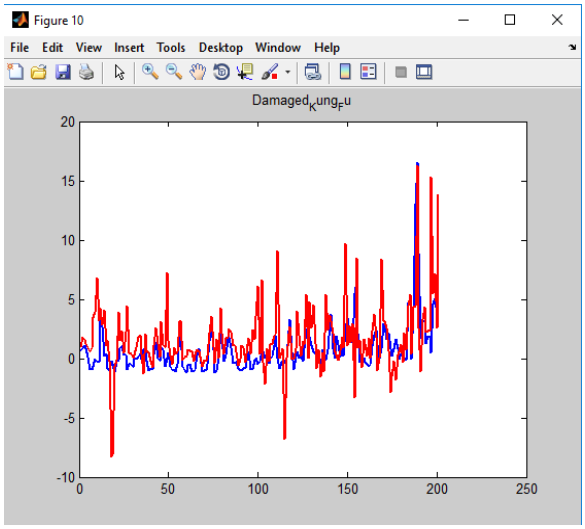


***Figura 27.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- First Bite

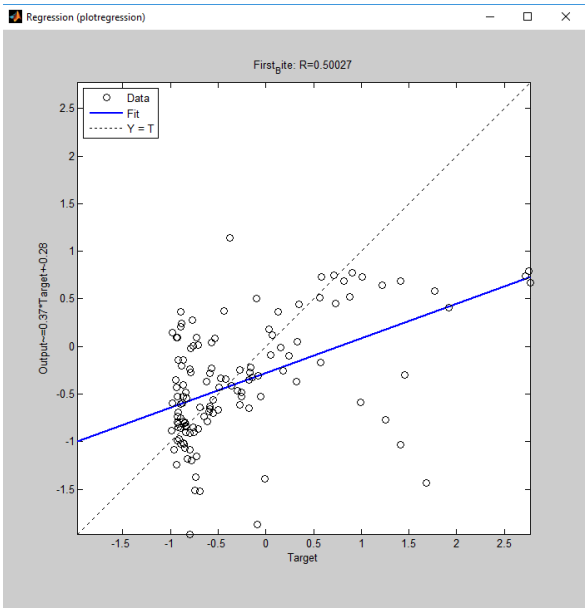


***Figura 28.*** *Regresión lineal de la película First Bite.*

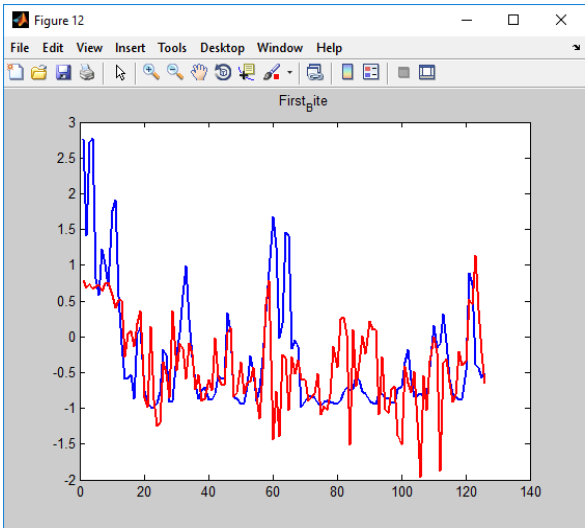


***Figura 29.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Full Service

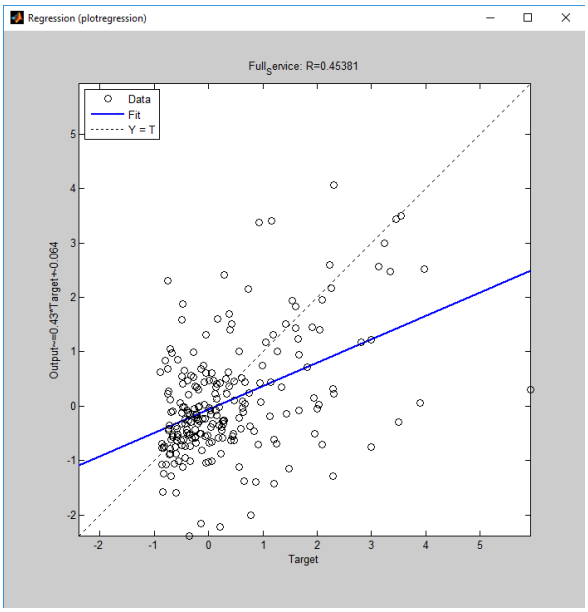


***Figura 30.*** *Regresión lineal de la película Full Service.*

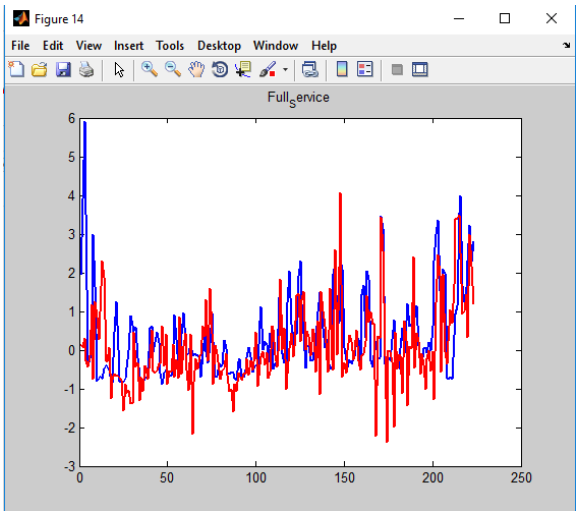


***Figura 31.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Islands

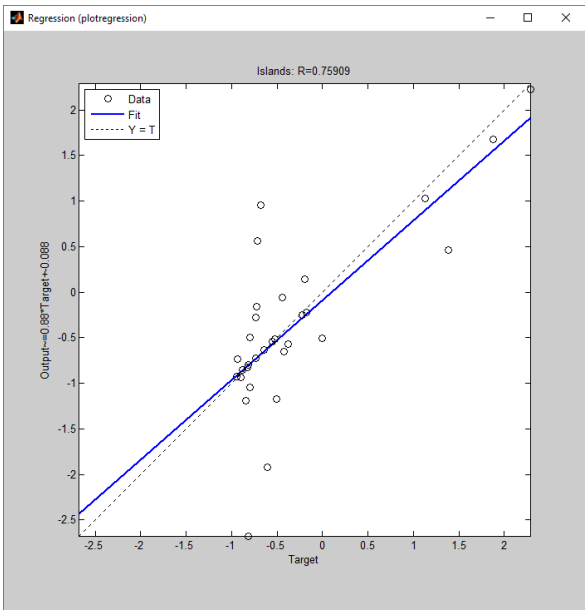


***Figura 32.*** *Regresión lineal de la película Islands.*

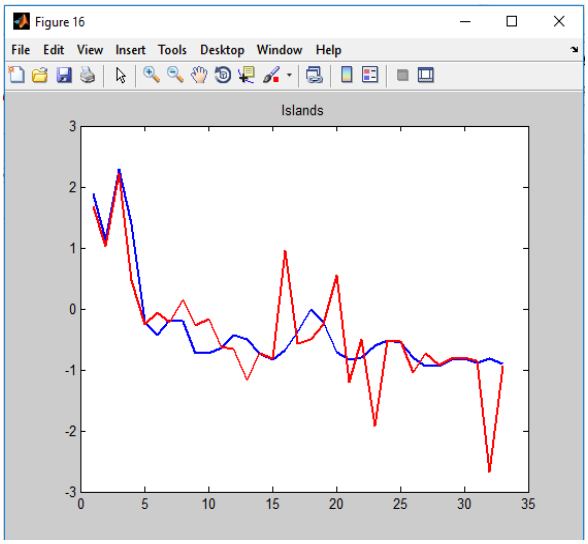


***Figura 33.*** *Comparativa entre los valores de salida del algoritmo (rojo) y los de ground truth (azul).*

- Lesson Learned

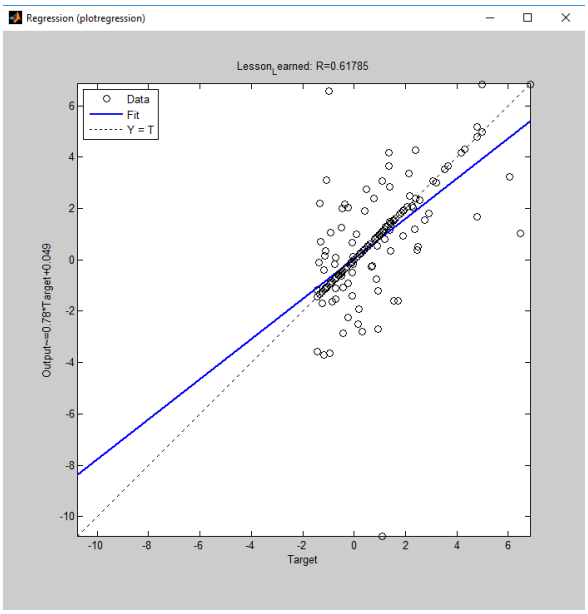


***Figura 34.*** *Regresión lineal de la película Lesson Learned.*

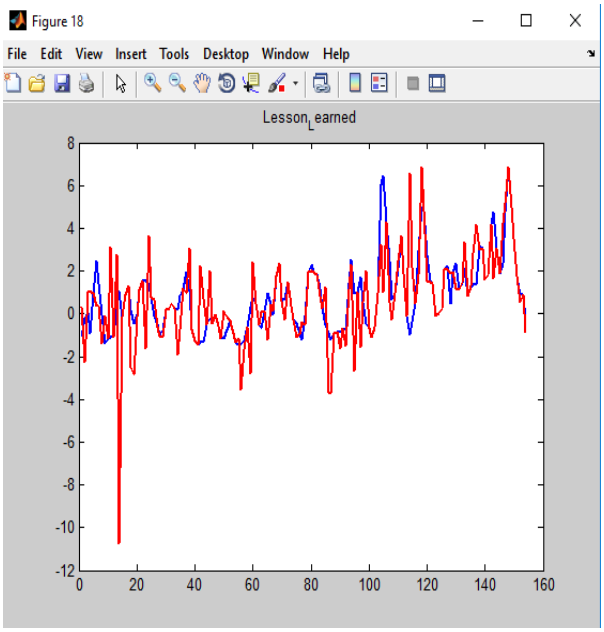


***Figura 35.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Norm

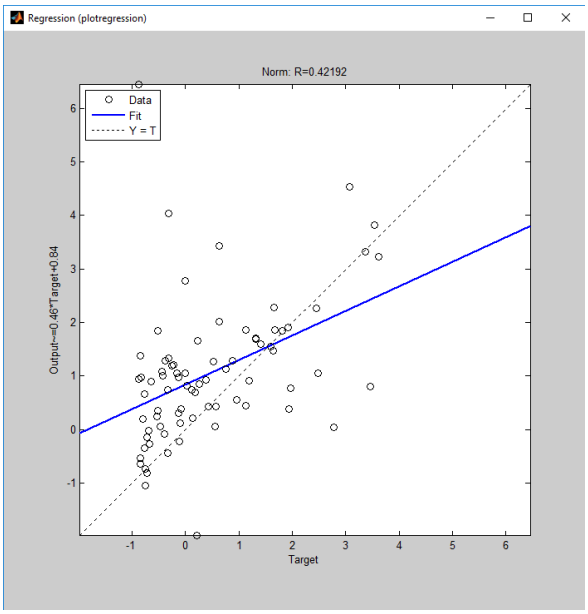


***Figura 36.*** *Regresión lineal de la película Norm.*

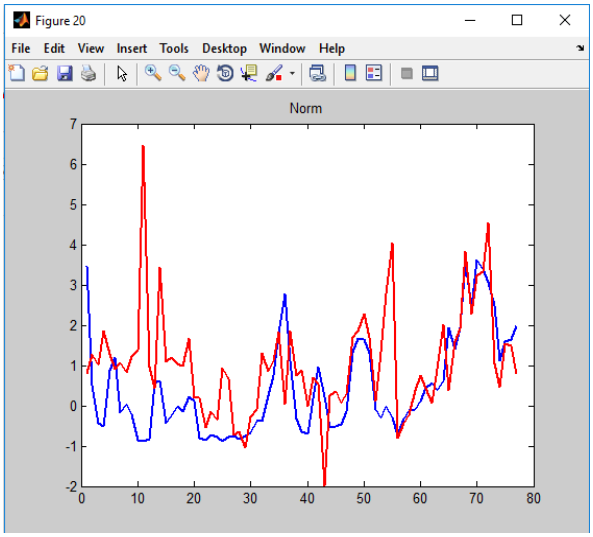


***Figura 37.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Nuclear Family

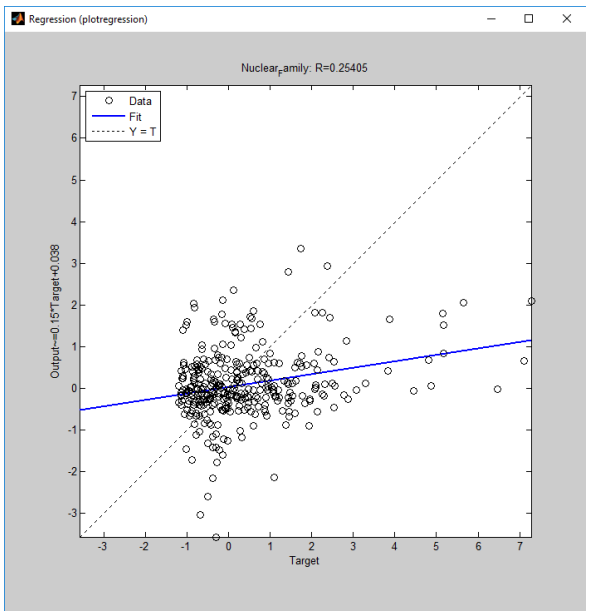


***Figura 38.*** *Regresión lineal de la película Nuclear Family.*

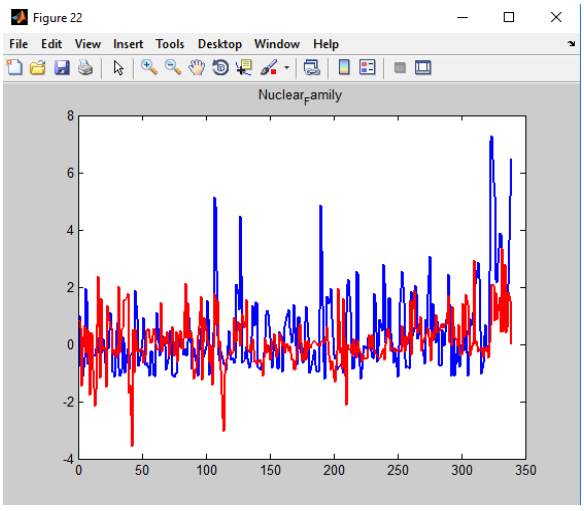


***Figura 39.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Riding The Rails

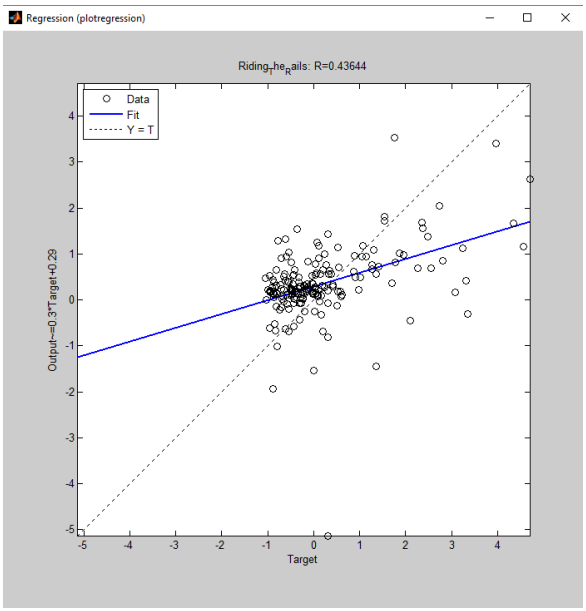


***Figura 40.*** *Regresión lineal de la película Riding the rails.*

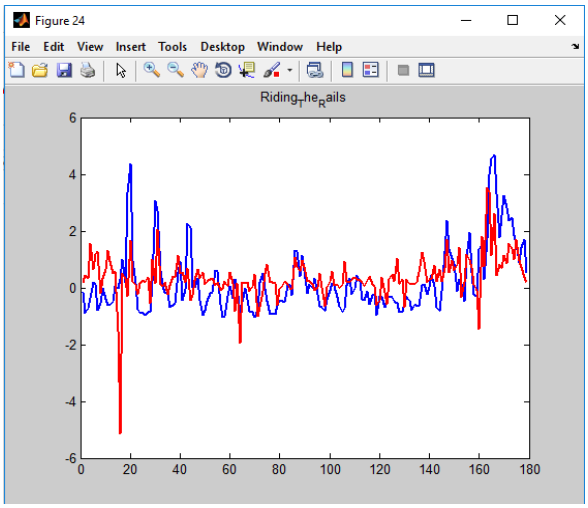


***Figura 41.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Sintel

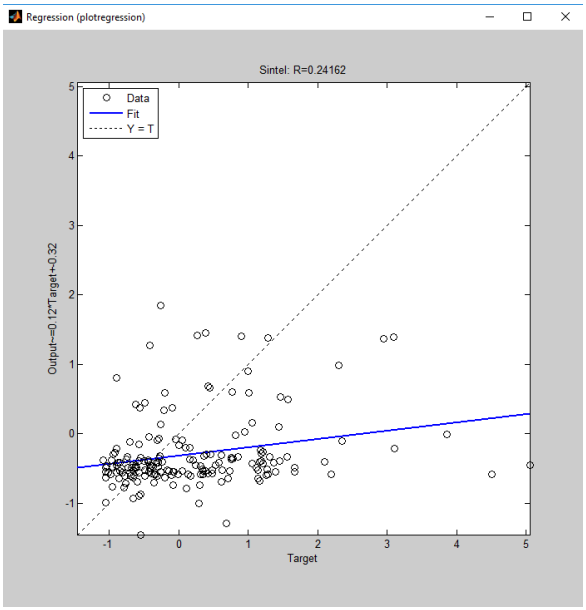


***Figura 42.*** *Regresión lineal de la película Sintel.*

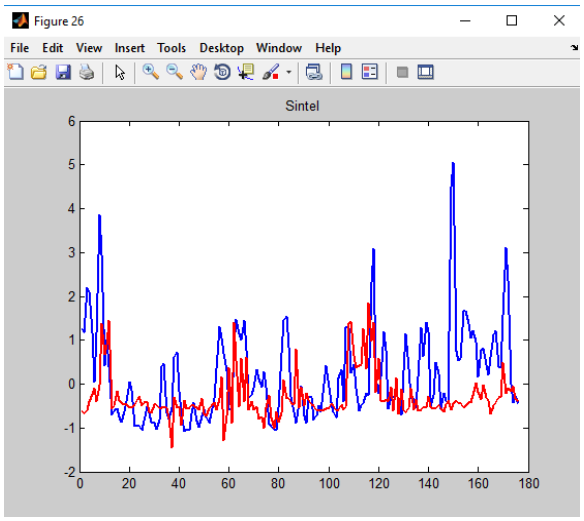


***Figura 43.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- Tears of Steel

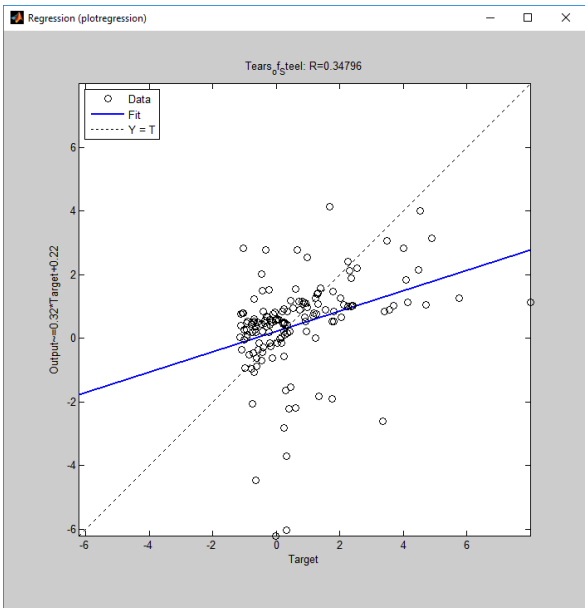


***Figura 44.*** *Regresión lineal de la película Tears of Steel.*

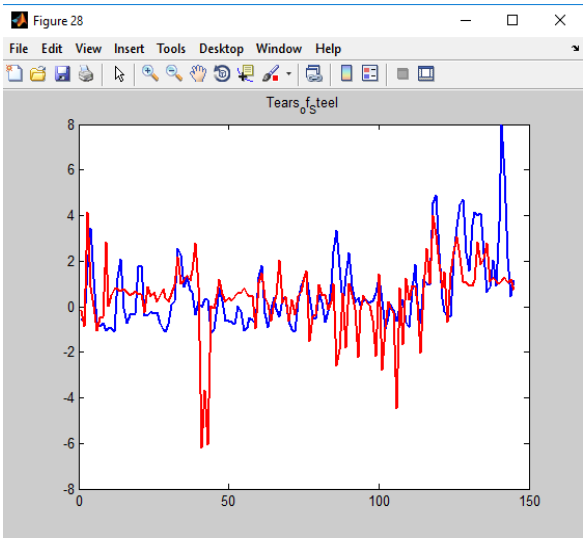


***Figura 45.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- The Room of Franz Kafka

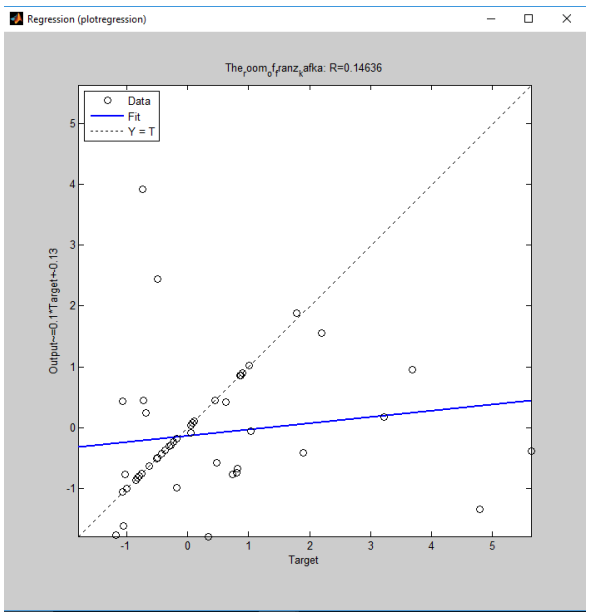


***Figura 46.*** *Regresión lineal de la película The room of Franz Kafka.*

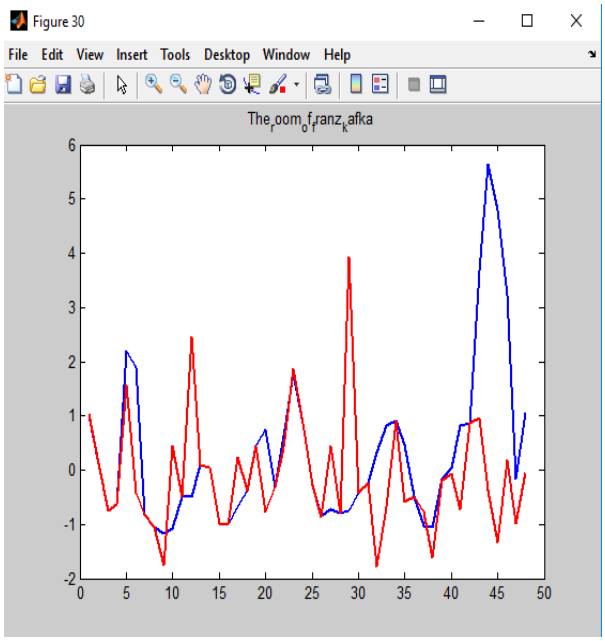


***Figura 47.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- The Secret Number

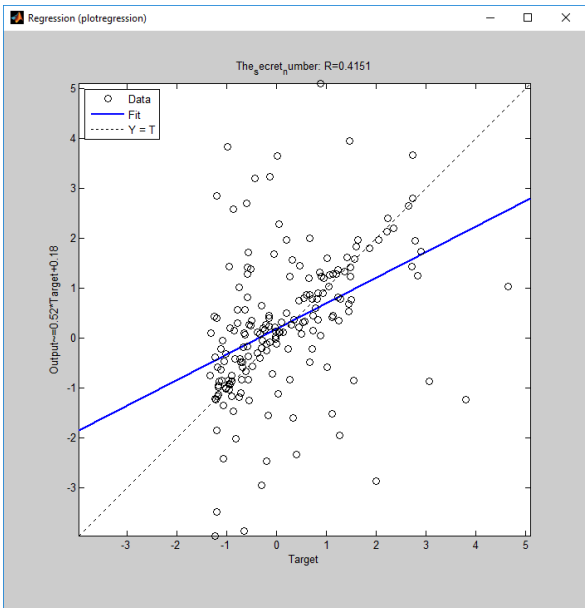


***Figura 48.*** *Regresión lineal de la película The secret number.*

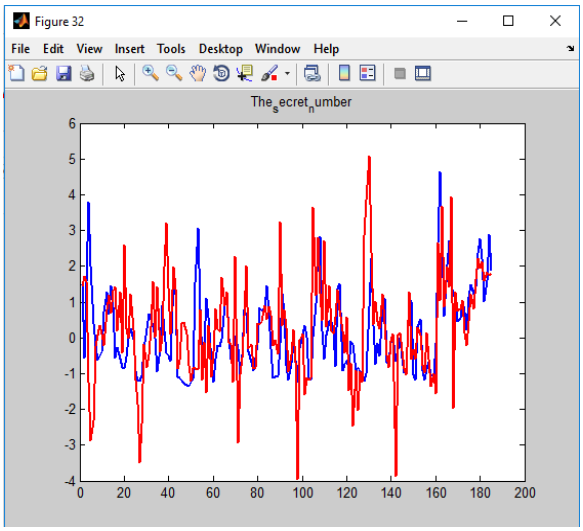


***Figura 49.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- To Claire From Sonny

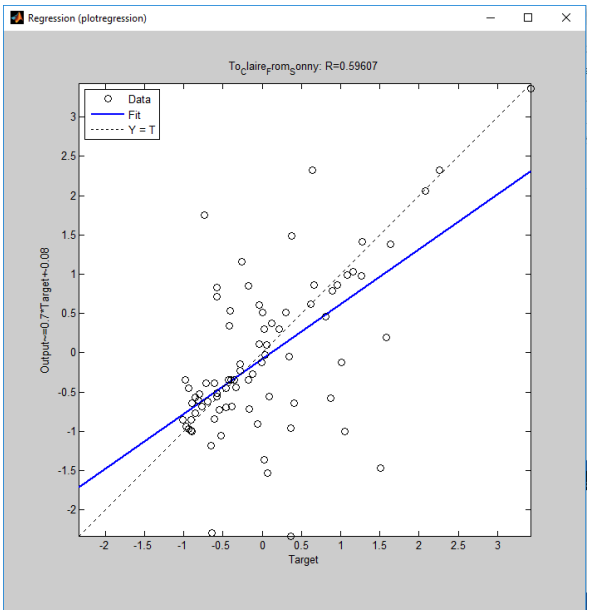


***Figura 50.*** *Regresión lineal de la película To Claire from Sony.*

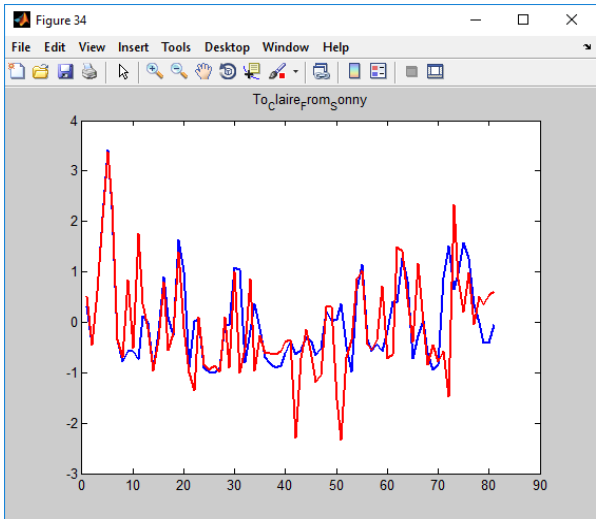


***Figura 51.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

- You Again

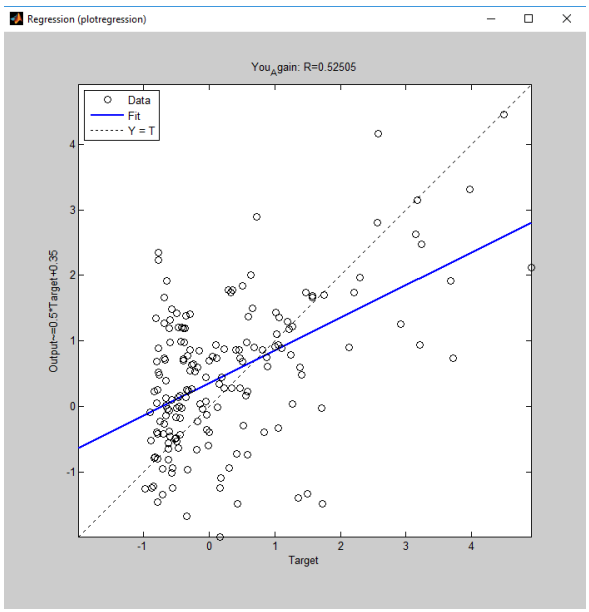


***Figura 52.*** *Regresión lineal de la película You again.*

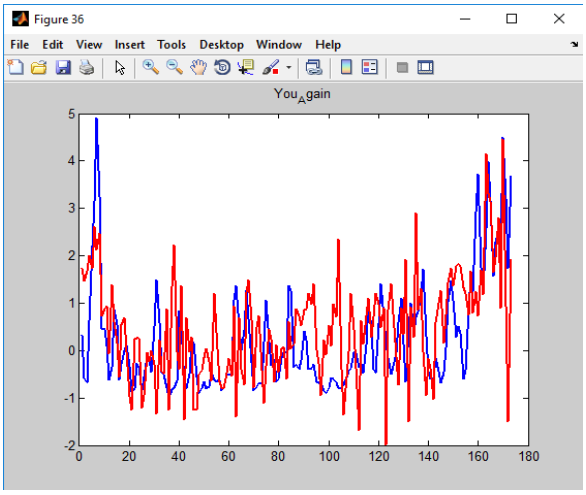


***Figura 53.*** *Comparativa entre los valores normalizados de salida del algoritmo (rojo) y los de ground truth (azul).*

## Referencias


[1] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret and Liming Chen, "*LIRIS-ACCEDE: A Video Database for Affective Content Analysis*".

[2] Mathworks, "*MATLAB, El lenguaje del cálculo técnico*". URL: http://es.mathworks.com/products/matlab/

[3] L. Itti, C. Koch, and E. Niebur, *"A Model of Saliency-Based Visual Attention for Rapid Scene Analysis"*. IEEE Trans. On PAMI, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[4] *C. Kayser, C. I. Petkov, M. Lippert, N. K. Logothetis, "Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map", Current Biology, vol. 15, no. 21, pp. 1943–1947, 2005.*

[5] O. Kalinli and S. S. Narayanan, *"A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech"*. Proc. of Interspeech, pp. 1941-1944, 2007.

[6] T. Tsuchida and G. W. Cottrell, *"Auditory saliency using natural statistics"*. Proceedings of the 34th Annual Conf. of the Cog Science Society pp. 1048-1053, 2012.

[7] B. Schauerte, R. Stiefelhagen, *"Wow! Bayesian Surprise for Salient Acoustic Event Detection"*. Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[8] J. Cadore, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Peláez-Moreno, *"Auditoryinspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement"*. Cog. Comp., no. 5, pp. 426–441, 2013.

[9] F. de-la-Calle-Silos, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Peláez Moreno, *"ASR Feature Extraction with Morphologically-Filtered Power-Normalized Cochleograms"*. Proc. of Interspeech, pp. 2430-2434, Singapore, September 2014.

[10] A. Gallardo-Antolín and J. M. Montero, *"Histogram Equalization-based Features for Speech, Music and Song Discrimination"*. IEEE Signal Proc Letters, vol. 17, no. 7, pp. 659-662, 2010.

[11] A. Gallardo-Antolín and R. San Segundo, *"UPM-UC3M System for Music and Speech Segmentation"*. VI Jornadas en Tecnología del Habla e Iberian SLTech Workshop (FALA 2010), pp. 421-424, 2010.

[12] Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez-Moreno, C., Valverde-Albacete, F.J. *"Feature extraction assessment for an acoustic-event classification task using the entropy triangle"* (2011) Proc. INTERSPEECH, pp. 309-312.

[13] J. Ludeña-Choez and A. Gallardo-Antolín, *"NMF-based Spectral Analysis for Acoustic Classification Tasks".* Advances in Nonlinear Speech Processing (NOLISP 2013), LNCS, vol. 7911, pp. 9-16, 2013.

[14] J. Ludeña-Choez and A. Gallardo-Antolín, *"NMF-based Temporal Feature Integration for Acoustic Event Classification".* Proc. Interspeech, pp. 2924-2928, 2013.

[15] J. Ludeña-Choez and A. Gallardo-Antolín, *"Feature Extraction Based on the High-Pass Filtering of Audio Signals for Acoustic Event Classification".* Comput. Speech and Language, (accepted), 2014.

[16] B. D. Coensel and D. Botteldooren, *"A model of saliency-based auditory attention to enviromental sound".* Proc. Int Congress on Acoustics (ICA), 2010.

[17] F. Tordini, A. S. Bregman, J. R. Cooperstock, A. Ankolekar and T. Sandholm, *"Toward An Improved Model Of Auditory Saliency".* Proc. ICAD, 2013.

[18] Ting Li, Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, Liming Chen, "*Continuous Arousal Self-assessments Validation Using Real-time Physiological Responses".*

[19] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, Liming Chen, "*From crowdsourced rankings to affective ratings".*

[20] Olivier Lartillot, Petri Toiviainen, Tuomas Eerola, "*A Matlab Toolbox for Music Information Retrieval*", in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2008.

[21] G. E. Hinton and R. R. Salakhutdinov*, "Reducing the dimensionality of data with neural networks."* Science vol. 313, no. 5786 (2006): 504-507.

[22] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, T.*, "A deep representation for invariance and music classification,"* ICASSP 2014, pp. 6984-6988, 2014.

[23] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik et al., *"Support vector regression machines,"* Advances in neural information processing systems, vol. 9, pp. 155–161, 1997.

[24] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, *"On the acoustics of emotion in audio: what speech, music, and sound have in common,"* Frontiers in psychology, vol. 4, pp. 1664–1078, 2013.

[25] G. E. Hinton and R. R. Salakhutdinov. *"Reducing the dimensionality of data with neural networks."* Science vol. 313, no. 5786 (2006): 504-507.

[26] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, *"The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data,"* in Affective Computing and Intelligent Interaction, 2007, vol. 4738, pp. 488–500.

[27] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, *"Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,"* Cognition & Emotion, vol. 24, no. 7, pp. 1153–1172, Nov. 2010.

[28] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *"DEAP: a database for emotion analysis using physiological signals,"* IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *"A multimodal database for affect recognition and implicit tagging,"* IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42–55,Jan. 2012.

[30] S. Carvalho, J. Leite, S. Galdo-A´ lvarez, and O. Gonc¸alves, *"The emotional movie database (EMDB): a self-report and psychophysiological study,"* Applied Psychophysiology and Biofeedback, vol. 37, no. 4, pp. 279–294, 2012.

[31] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, *"A benchmarking campaign for the multimodal detection of violent scenes in movies,"* in Proceedings of the 12th International Conference on Computer Vision, ser. ECCV'12, 2012, pp. 416–425.

[32] M. Horvat, S. Popovic, and K. Cosic, *"Multimedia stimuli databases usage patterns: a survey report,"* in Proceedings of the 36nd International ICT Convention MIPRO, 2013, pp. 993–997.

[33] K. Sun, J. Yu, Y. Huang, and X. Hu, *"An improved valencearousal emotion space for video affective content representation and recognition,"* in IEEE International Conference on Multimedia and Expo, 2009, pp. 566–569.

[34] K. Kim, K.-H. Lin, D. B. Walther, M. A. Hasegawa-Johnson and T. S. Huang, *"Automatic detection of auditory salience with optimized linear filters derived from human annotation",* Pat Recogn Lett, vol. 38, pp. 78-85, 2014.

[35] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, *"Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention,"* IEEE Trans. Multimedia, vol.15 (7), p.1553-1568, 2013.